University of Bath

PHD

Carbohydrate-degrading enzymes from the thermophilic ethanologen Geobacillus thermoglucosidasius

Espina Silva, Giannina

Award date:
2015

Awarding institution:
University of Bath

Link to publication

Alternative formats
If you require this document in an alternative format, please contact:
openaccess@bath.ac.uk

Copyright of this thesis rests with the author. Access is subject to the above licence, if given. If no licence is specified above, original content in this thesis is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC-ND 4.0) Licence (https://creativecommons.org/licenses/by-nc-nd/4.0/). Any third-party copyright material present remains the property of its respective owner(s) and is licensed under its existing terms.

Take down policy
If you consider content within Bath's Research Portal to be in breach of UK law, please contact: openaccess@bath.ac.uk with the details. Your claim will be investigated and, where appropriate, the item will be removed from public view as soon as possible.

Download date: 05. May. 2023

Carbohydrate-degrading enzymes from the thermophilic ethanologen *Geobacillus thermoglucosidasius*

Giannina Angelica Espina Silva

A thesis submitted for the degree of Doctor of Philosophy

University of Bath

Department of Biology and Biochemistry

February 2015

COPYRIGHT

Attention is drawn to the fact that copyright of this thesis rests with the author. A copy of this thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that they must not copy it or use material from it except as permitted by law or with the consent of the author.

This thesis may be made available for consultation within the University Library and may be photocopied or lent to other libraries for the purposes of consultation.

Signed on behalf of the Faculty of Science

ACKNOWLEDGMENTS

On the completion of my PhD I would like to acknowledge the people who have contributed to it in one way or another. First of all, I must extend heartfelt thanks to my supervisors Professor Michael Danson and Dr Susan Crennell for their trust and for giving me the incredible opportunity to carry out research under their expert supervision. I am indebted for their teaching, encouragement, help, guidance, constant support and generous amount of time. I really appreciate that they were always willing to be available for me and for their patience, kindness and care.

I would also like to express my gratitude to everyone in the Centre for Extremophile Research and both past and present members from the Labs 1.33 and 1.28, who made days pass very quickly! I really appreciate their friendship and enthusiasm as well as their assistance and suggestions to my project; it was a pleasure to work with them. Special thanks go to Dr Jon Extance, Dr Nia Marrot, Dr Chris Hills, Chris Vennard, Dr David Hough and Dr Momna Hejmadi who really helped me when I started my research; Ali Hussein, Luke Williams, Lisa Buddrus who helped me when I was finishing, and Carolyn Williamson who constantly helped me throughout my PhD. I would also like to recognise the advice I received from the postdocs Dr Jeremy Bartosiak-Jentys, Dr Steven Bowden, Dr Gareth Cooper, Dr Christopher Ibenegbu, Dr Charlie Bennett and Dr Leann Bacon whose expertise and encouragement helped me a lot when planning new experiments.

I thank Dr Rob Meijers, Dr Stephane Boivin, Dr Guillaume Pompidor and Dr. Thomas Schneider at the High Throughput Crystallization Facility of the EMBL-Hamburg for their help with the analysis of my samples during the 8th Protein Expression, Purification and Characterisation, EMBO practical course (PEPC8) and my follow-up visit to PETRA III Synchrotron at DESY kindly funded by P-CUBE.

I am also very grateful to my country and especially to the National Commission for Scientific and Technological Research (CONICYT) of Chile for funding my PhD studies at the University of Bath through a *Beca Chile* grant. Financial contributions from TMO Renewables Ltd is also recognised and appreciated, along with the travel grants from the Biochemical Society and the Society for General Microbiology that allowed me to present my research in the 36th Symposium on Biotechnology for Fuels and Chemicals, Florida, US.

On a personal note, I am very grateful for the love and encouragement that I have received from my parents, Maria Angelica Silva and Sergio Espina, and my brothers, Sergio and Harold, who despite the distance during the past four years, have always been there for me and have never wavered in their belief in what I can achieve. I would also like to extend my thanks to my friends in Chile, especially Ximena Fernández, Daniela Neira, Fabiola Morales and Fabiola Altimira and also my friends in the UK, especially the families Castillo-Escribano and Leyton-Muñoz, Jocelyn Dunstan, Dan and Paula Knowlson and Simon Watkins who have always supported me and made my time here really enjoyable.

Last but not least, my very special thanks go to my wonderful husband Carlo Cortés, to whom I dedicate this thesis. His love, support, patience and understanding have been essential in accomplishing this; I am so grateful for his courage that made him marry me in the first place, and then move with me to another continent so far away from our home country and families. It has not been easy but it surely has been the most enriching and unforgettable experience of our lives and I am so happy that we made it together.

TABLE OF CONTENTS

ABSTRACT

It is widely known that fossil fuels are limited; consequently, the generation of new sources of energy in a clean and environmentally friendly manner is a research priority. Bioethanol appears to be one potential solution, especially second-generation production from renewable biomass.

In order to use lignocellulosic feedstock to produce bioethanol, its polysaccharide components, cellulose and hemicellulose, must be hydrolysed into soluble sugars, which can then be converted into ethanol by fermentative microorganisms such as *Geobacillus thermoglucosidasius* TM242 used by the company ReBio Technologies Ltd.

To date, the cost of commercial enzymes used during the hydrolysis process remains a major economic consideration in the production of second-generation bioethanol as an alternative fuel. The research project presented in this thesis aims to improve this rate-limiting step of microbial bioethanol production through an investigation of the different enzymes associated with hemicellulose hydrolysis.

Firstly, the TM242 genome sequence revealed a number of genes encoding glycoside-hydrolases. Six of these genes were cloned and expressed in *E. coli* and the recombinant enzymes characterised; three of them, two β-xylosidases and an α-arabinofuranosidase, are relevant to xylan hydrolysis, and were found to be highly active and thermostable. Crystallisation of one of the β-xylosidases permitted the determination of a high-resolution (1.7 Å) structure of the apo-enzyme along with a lower resolution (2.6 Å) structure of the enzyme-substrate complex, resulting in the first reported structure of a GH52 family member (Espina *et al.,* 2014).

Secondly, as the TM242 microorganism lacks xylanase enzymes, four genes encoding xylanases from closely-related *Geobacillus* strains were cloned and expressed in *E. coli,* with one of them being also successfully cloned and expressed in *G. thermoglucosidasius* TM242. This heterologous xylanase was secreted in active form representing an enhanced biomass utilisation by TM242.

In conclusion, it is felt that the findings presented here have the potential to make a valuable contribution towards second-generation bioethanol production.

LIST OF ABBREVIATIONS

$\|F_{hkl}\|$	Structure factor amplitude of the reflection *hkl,*
<I/σ(I)>	Average signal-to-noise ratio
2SPYNG	Soy Peptone Yeast Extract medium
A	Absorbance
ABC	ATP-binding cassette
AbfA	GH51 α-L-arabinofuranoside from *G. stearothermophilus* T-6
AFEX	Ammonia fibre explosion
AldDH	acetaldehyde dehydrogenase
APS	Ammonium persulphate
***B* factor**	Temperature factor
Bis-Tris	2-Bis(2-hydroxyethyl)amino-2-(hydroxymethyl)-1,3-propanediol
BLAST	Basic Local Alignment Search Tool
BSA	Bovine Serum Albumin
C5	Pentose sugars
C6	Hexose sugars
CAMR	Centre for Applied Microbiological Research, Porton Down
CATH	Hierarchic classification of protein domain structures bassed on Class, Architecture, Topology and Homologous superfamily
CAZy	Carbohydrate-Active enZYmes Database
CBM	Carbohydrate Binding Modules
CBP	Consolidated bioprocessing
CCP4	Collaborative Computational Project Number 4
CDD	Conserved Domain Database
CoA	Coenzyme A
COOT	Crystallographic Object-Oriented Toolkit
CV	Column volume
DESY	Deutsches Elektronen-Synchrotron
DNS	3 5-dinitrosalicylic acid
dNTP	Deoxyribonucleotide
EBI	European Bioinformatics Institute
EDTA	Ethylenediaminetetraacetic acid
EMBL	European Molecular Biology Laboratory
EMBO	European Molecular Biology Organisation
ESI	Electrospray ionisation
ExPASy	Expert Protein Analysis System
F_{calc}	Calculated structure factor
F_{hkl}	Structure factor of the reflection *hkl,*
F_{obs}	Experimentally obtained structure factor
FPLC	Fast protein liquid chromatography
GH	Glycoside Hydrolase
GSxyn	GH52 β-xylosidase from *G. stearothermophilus* 1A05585
HEPES	2-[4-(2-hydroxyethyl)piperazin-1-yl]ethanesulfonic acid
HF	High-Fidelity
HiAXHd3	α-arabinofuranosidase from *Humicola insolens*
hkl	Scattering plane (with Miller Indices hkl)
HMF	5-(Hydroxymethyl)furfural
HPLC	High-performance liquid chromatography
HTX	High Throughput Crystallization Facility
HUS	Hemicellulose Utilisation System
I_{hkl}	Intensities of the waves diffracted from the crystal plane hkl

IMAC	Immobilised-Metal Affinity Chromatography
IPCC	Intergovernmental Panel on Climate Change
IPTG	Isopropyl β-D-1-thiogalactopyranoside
ITC	Isothermal Titration Calorimetry
IXC56	Intracellular xylanase from *G. thermoglucosidasius* C56-YS93
IXT6	Intracellular xylanase from *G. stearothermophilus* T-6
Kav	Partition coefficient
k_{cat}	Catalytic constant that describes the turnover rate
k_{cat}/K_M	Specificity constant that denotes the catalytic efficiency of an enzyme
keV	Kiloelectronvolt
K_i	Dissociation constant of an inhibitor
K_M	Michaelis constant
LB	Lysogeny Broth
LCMS	Liquid Chromatography-Mass Spectrometry
ldh	Lactate dehydrogenase
m/z	Mass-to-charge ratio
MAD	Multiple Wavelength Anomalous Diffraction/Dispersion
MDL	Molecular Dimensions Limited
MES	2-(N-morpholino)ethanesulfonic acid
MilliQ	Ultrapure water
MIR	Multiple Isomorphous Replacement
MOPS	3-(N-morpholino)propanesulfonic acid
M_r	Relative molecular mass
NagZ	GH3 β-N-hexosaminidase from *Bacillus subtilis*
Ni-NTA	Nickel-nitrilotriacetic
NMR	Nuclear Magnetic Resonance
NREL	National Renewable Energy Laboratory (US)
OD	Optical density
PCR	Polymerase Chain Reaction
PCT	Pre-crystallisation test
PDB	Protein Data Bank
pdh	Pyruvate dehydrogenase
PDHC	Pyruvate dehydrogenase complex
PEG	Polyethylene glycol
PEPC8	8th Protein Expresion, Purification and Characterisation, practical course
pfl	Pyruvate formate lyase
PHENIX	Python-based Hierarchical ENvironment for Integrated Xtallography
pI	Isoelectric point
PISA	Protein Interfaces, Surfaces, and Assemblies
***p*NP**	Para-nitrophenol
***p*NP-AF**	*p*-Nitrophenyl-α-L-arabinofuranoside
***p*NP-AP**	*p*-Nitrophenyl-α-L-arabinopyranoside
***p*NP-GP**	*p*-Nitrophenyl-β-D-glucopyranoside
***p*NP-NAG**	*p*-Nitrophenyl N-acetyl-β-D-glucosaminide
***p*NP-XP**	*p*-Nitrophenyl-β-D-xylopyranoside
PQS	Probable Quaternary Structure
PUp2n38	Uracil phosphoribosyltransferase constitutive promoter
Q-TOF	Quadrupole-Time-of-Flight
r.m.s.d	Root Mean Square Deviation from ideal geometry
R_{cryst}	Residual factor or reliability factor
Rf	Relative mobility
R_{free}	Analogous to the normal Rcryst, but for only 5-10% of the dataset
SAD	Single Wavelength Anomalous Diffraction/Dispersion
SANS	Small Angle Neutron Scattering

SAP	Shrimp Alkaline Phosphatase
SAXS	Small Angle X-ray Scattering
SDS-PAGE	Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis
SEC	Size-Exclusion Chromatography
SeMet	Selenomethionine
SIR	Single Isomorphous Replacement
SOC	Super Optimal with Catabolite repression medium
SPy1599	GH1 6-phospho-β-D-glucosidase from *Streptococcus pyogenes*
TAE	Tris-acetate-EDTA buffer
TB	Terrific Broth
TCEP	Tris-(2-carboxyethyl)phosphine
TEMED	Tetramethylethylenediamine
TLC	Thin Layer Chromatography
$\mathbf{T_m}$	Melting temperature
tm1225	Predicted glycoside hydrolase from *Thermotoga maritima* msb8
TRIS	2-Amino-2-hydroxymethyl-propane-1,3-diol,
U	Unit of enzyme activity (µmol/min)
USM	Modified Urea Salts medium
V	Volt
$\boldsymbol{V_{max}}$	Maximum velocity of reaction
X12739	Extracellular xylanase from CAMR12739
X5420	Extracellular xylanase from CAMR5420
XC56	Extracellular xylanase from *G. thermoglucosidasius* C56-YS93
XDH	Xylose dehydrogenase
X-GAL	5-bromo-4-chloro-3-indolyl β-D-galactopyranoside
XT6	Extracellular xylanase from *G. stearothermophilus* T-6
XynB2	GH52 β-xylosidase from *G. stearothermophilus* T-6
XynB3	GH43 β-xylosidase from *G. stearothermophilus* T-6
YENB	Yeast Extract Nutrient Broth
$\boldsymbol{\alpha_{hkl}}$	Phase angle of the diffracted beam
$\mathbf{\Delta G^{diss}}$	Change in free energy of assembly dissociation
$\mathbf{\Delta G^{int}}$	Change in the solvation free energy of assembly.formation
$\mathbf{\Delta H_a}$	Binding enthalpy
$\boldsymbol{\varepsilon_M}$	Molar absorption coefficient
$\boldsymbol{\rho_{xyz}}$	Electron density at every position *x,y*, *z*

CHAPTER 1: General Introduction

1.1. World Energy Crisis

Since the beginning of the industrial age in the 18th century, fossil fuels such as coal, crude oil and natural gas have been humanity's most important source of energy. It has been estimated by the BP Statistical Review of World Energy (2014) that their wide use contributes approximately 87% of the total energy consumption in the world and that, due to industrial development and population growth, the global demand continues to increase, leading us to a strong dependency on this finite resource.

Since the 1970s, it has been recognized that a world energy crisis is facing our society as we have continuously exploited the limited fossil fuel sources available on earth, which after reaching their extraction peaks will enter terminal decline. According to the BP Energy Outlook (2014) it is projected that by 2035 the global energy consumption will rise by 41% in comparison to 2012, which, if no action is taken, will obviously threaten the international energy security system and geopolitical stability.

In addition, the excessive use of fossil fuels and their combustion have been causing substantial environmental damage, such as pollution, acid rain and global warming, which has been described to be primarily associated to the continuous and massive emissions of greenhouse gases (*e.g.* CO_2, CH_4 and N_2O) that have reached atmospheric levels unprecedented in at least the last 800,000 years (Liu *et al.*, 2007; IPCC, 2014). Since the 1950s, climate changes linked to warming of the atmosphere and ocean have been observed, including an increase in cold temperature extremes, an increase in warm temperature extremes, diminished amounts of snow and ice, high sea levels and an increase in the number of heavy precipitation events in a number of regions (IPCC, 2014).

These issues have started to seriously affect our society, and the spiralling cost of energy, both financially and environmentally, has raised serious concerns with the public and governments across the world. Consequently, in response to the energy crisis alert and immediate environmental deterioration associated with conventional non-renewable fuel usage, there is a strong drive to find viable alternatives to the use of fossil fuels.

Some long-term alternatives include the utilisation of energy sources such as wind, hydro, geothermal, solar and biomass, all of which are renewable and have no direct carbon dioxide emissions (Liu *et al.*, 2007). Biomass energy is a special case as, in spite of releasing carbon with a carbon-to-energy ratio similar to that of coal, it has already absorbed an equal amount of carbon from the atmosphere prior to its emission, so that net emissions of carbon from biomass fuels are zero over its life cycle (Cherubini & Strømman, 2011). This offers a significant saving in the greenhouse gases emissions when compared to fossil fuels, since, even though they are also derived from the remains of organic matter, their combustion returns carbon that was sequestered hundreds of millions of years ago. Furthermore, biomass is the most significant source of renewable carbon in the biosphere and it is deemed to be one of the most auspicious renewable energy alternatives available in the short-term.

1.2. Biofuels

The term biofuel refers to solid (*e.g.* biochar), liquid (*e.g.* bioethanol, vegetable oil and biodiesel) or gaseous (*e.g.* biogas, biosyngas and biohydrogen) fuels that are predominantly produced from plant-based material and present a renewable energy source derived from the solar energy through photosynthesis (Demirbas, 2009). A wide range of biomass feedstock can be used to produce biofuels and this can be done either by thermal, chemical or biochemical conversion (Menon & Rao, 2012) and, according to the feedstock from which they are produced, biofuels can be classified into first generation (produced directly from food crops), second generation (produced from lignocellulose or biowaste) and third generation (produced from algae) (Lee & Lavoie, 2013).

The global production and use of biofuels have increased dramatically in the past few years, mainly supported by the fact that many countries have set ambitious goals to substitute with liquid biofuels a part of the fossil fuels used for transportation. This includes those countries of the European Union that aim to reach a quota of 10% biofuels in the transport sector by 2020 (The European Parliament and the Council of the European Union, 2009) and the US Department of Energy Office with a scenario for supplying 30% of the 2004 petrol demand with biofuels by the year 2030 (NREL, 2007). Even though transport is not the only area of heavy fossil fuel usage, it constitutes about 34% of the total world energy consumption and is responsible for at least 23% of global CO_2 emissions, with a steadily increasing demand (Morales *et al.*, 2015; Rabinovitch-Deere *et al.*, 2013) . Therefore, energy for transportation might

represent the easiest form where conversion to more sustainable fuels can provide an interim solution.

Bioethanol is the most widely used liquid biofuel, with the US, Brazil, and China as the largest worldwide bioethanol producers. Due to its higher heat of vaporization and octane number, it can be mixed to 5% with petrol in conventional vehicles with no modifications (currently a requirement of the Renewable Transport Fuel Obligation (RTFO) in the UK (Department for Transport, 2007)), blended up to 85% (E85) in flex-fuel vehicles, or used as the sole fuel in dedicated engines (Haghighi Mood *et al.*, 2013). When mixed with petrol, the oxygen content of bioethanol results in a greater extent of hydrocarbon oxidation and thus diminishes the emissions of greenhouse gases (Sánchez & Cardona, 2008).

These characteristics have encouraged a substantial growth in the bioethanol global market, which increased from less than a billion litres in 1975, to more than 50 billion litres in 2006, and it is predicted that global ethanol production will expand to reach 158 billion liters by 2023 (OECD-FAO, 2014). Currently, most ethanol produced in the world is first-generation from agricultural crops that can be easily hydrolysed into soluble sugars and then fermented into ethanol; sugar cane is the preferred raw material for ethanol production in Brazil, India, and South Africa, whereas corn is used in the US and sugar beet in France (Zaldivar *et al.*, 2001). However, challenges such as capacity limitations (feedstock availability and supply), food *versus* fuel issues, high feedstock prices, land and fresh water use, and intensive agricultural contributions, have led to investigations of second-generation bioethanol that can overcome some of these obstacles and help to supply the growing demand for sustainable renewable energy sources (Kumar & Murthy, 2011).

1.3. Second generation bioethanol production

In contrast to first-generation, second-generation bioethanol can be produced from various non expensive lignocellulosic feedstock like agricultural and forestry residues (*e.g.* straws, hulls, stems and stalks), municipal and industrial wastes that largely consist of plant materials and specific bioenergy crops (*e.g.* grasses such as sorghum, miscanthus and switchgrass, and fast-growing trees such as willow and poplar grown on waste land (Simmons *et al.*, 2008)). This represents a much better and environmentally acceptable alternative in terms of long term sustainability as lignocellulose is the most abundant renewable natural resource on earth. In addition,

being an available substrate for conversion to biofuels, land conversion is not required and therefore, its use will not impact the provision of food for humans or of feed for livestock (Lal, 2008; Sheehan, 2009). Furthermore due to its abundance and low cost it has been suggested that the potential quantity of bioethanol that could be produced from it can be over an order of magnitude larger than that produced from food crops (Demain *et al.*, 2005).

1.3.1. Composition of the lignocellulose

Lignocellulosic biomass is mainly composed of a high cellulose content, typically in the range of approximately 35 to 50%, embedded in a matrix of other structural biopolymers, primarily hemicelluloses and lignin, which comprise 20 to 35% and 10 to 25% respectively (Saha, 2003).

Cellulose is a carbohydrate homopolymer synthesised in nature as individual molecules (linear chains of β-1-4 linked glucosyl residues), approximately 30 of which undergo self-assembly into larger units known as elementary fibrils (protofibrils), which are packed into larger units called microfibrils as illustrated in Figure 1.1 (Arantes & Saddler, 2010). These are unbranched fibrils composed of approximately 30-36 glucan chains aggregated laterally by means of non-covalent interactions (hydrogen bonding and Van der Waals forces) to produce a stable crystalline lattice structure insoluble in water and most organic solvents (Arantes & Saddler, 2010; Lynd *et al.*, 2002).

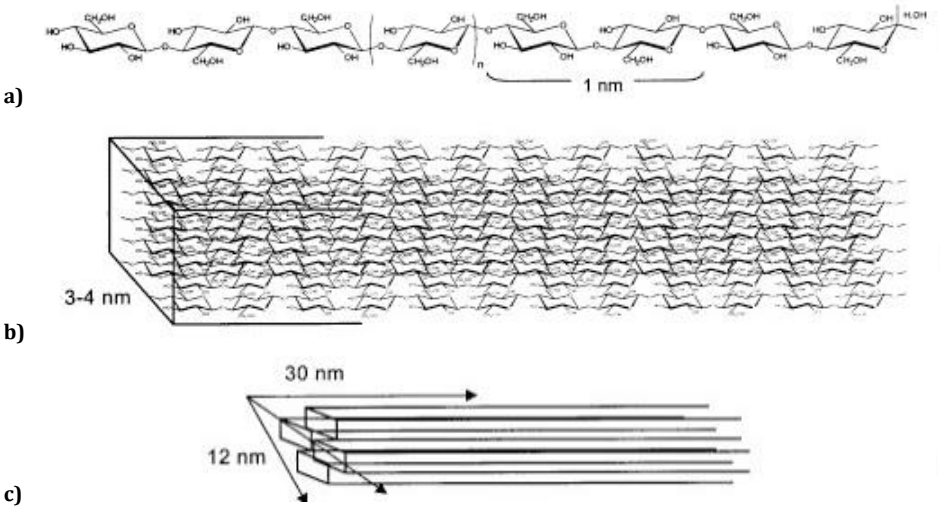


Figure 1.1. Structure and association of cellulose. (a) Individual cellulose chains, with an indication of the length of the basic structural unit, cellobiose. (b) Framework of cellulose chains in the elementary fibril. (c) Microfibril. Image obtained from Sun, 2010.

Hemicellulose is a highly-branched carbohydrate heteropolymer composed of pentoses (*e.g.* D-xylose, L-arabinose), hexoses (*e.g.* D-galactose, L-galactose, D-mannose, L-rhamnose, L-fucose), and/or uronic acids (*e.g.* D-glucuronic, D-4-O-methylgalacturonic and D-galacturonic acids) (Petrova & Ivanova, 2014). There are many different hemicelluloses characterised by β-(1,4)-linked backbones of xylose, mannose, or glucose, that can be further classified according their substitutions (Scheller & Ulvskov, 2010). Examples include xylans (characterised by a backbone of β-(1,4)-linked xylose residues), mannan (with a backbone of β-(1,4)-linked mannose), and glucomannans (with a backbone of β-(1,4)-linked mannose and glucose).

The exact composition and structure (degree of polymerisation, length of the backbone and the nature and degree of substitution) of hemicelluloses varies strongly among different plant species, tissues and cell types. For example, cereal grains commonly contain mostly arabinoxylans (composed of a xylan backbone largely substituted with D-arabinose, shown in Figure 1.2a), whereas glucuronoxylans (composed of a xylan backbone largely substituted with 4-O-methyl-D-glucuronic acid, shown in Figure 1.2b) are found in hardwoods, and softwood mostly contains glucomannans (shown in Figure 1.2c) (Pollet *et al.*, 2010; Saha *et al.*, 2009; Scheller & Ulvskov, 2010).

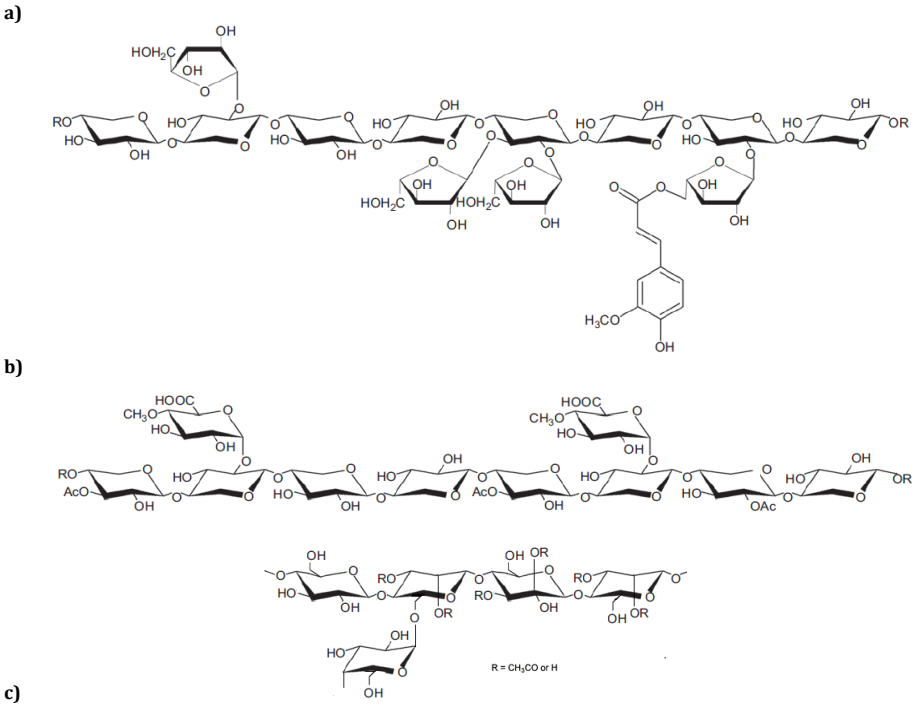


Figure 1.2. Structure of three different hemicelluloses. (a) Arabinoxylan from cereal grains. (b) Glucuronorabinoxylan from hardwoods. (c) Glucomannan from softwoods. Images a and b obtained from Pollet *et al.*, 2010.

On the other hand, lignin is a complex aromatic polymer (Figure 1.3) synthesised from the phenol dehydration of three monomeric alcohols (lignols) derived from *p*-cinnamic acid: *trans*-*p*-coumaryl, *trans*-*p*-coniferyl, and *trans*-*p*-sinapyl alcohol. Its exact composition varies from species to species, and the degree of cross-linking possible in lignin is dependent on the quantity of water and polysaccharides, as well as molecular weight and chemical functionalisation (Ghaffar & Fan, 2013).

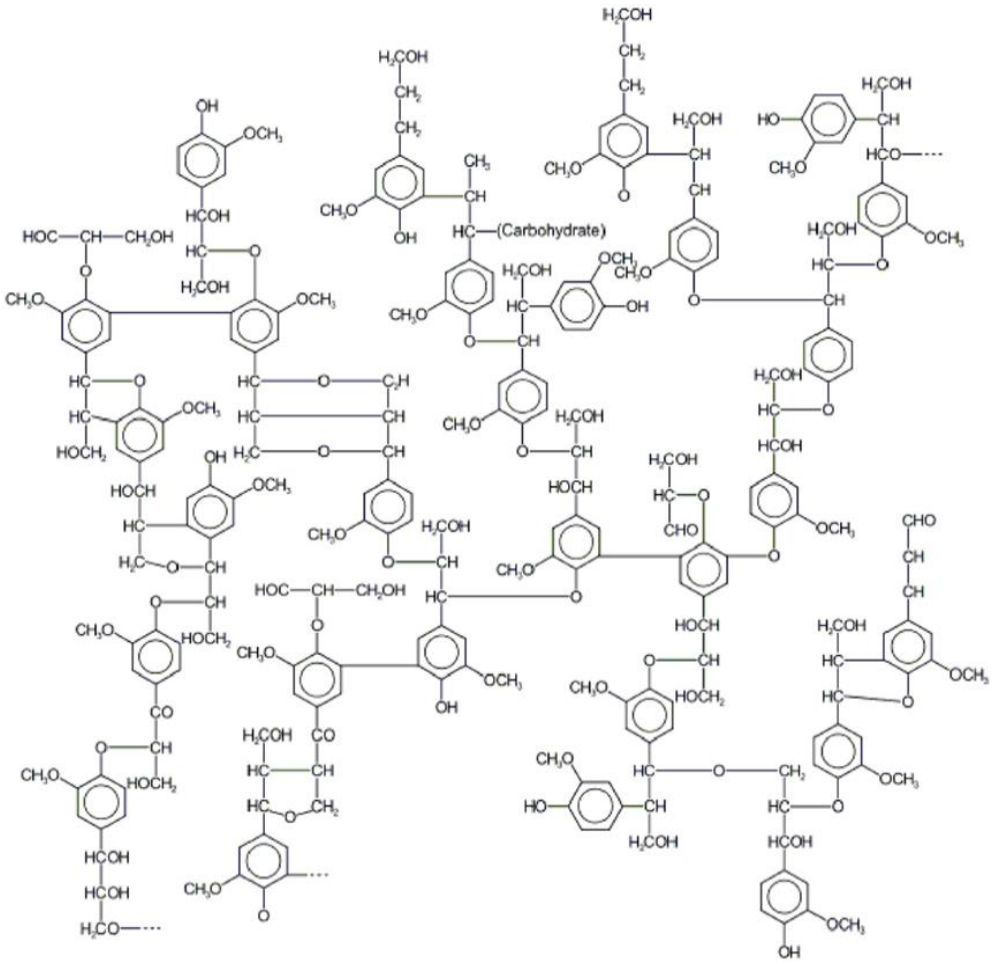


Figure 1.3. Representative lignin structure. Image obtained from Christopher *et al.*, 2014.

The plant cell wall is clearly important as it provides support and shape for the plants, allowing them to stand upright and endure large physical forces, and is also a barrier against the environment and potentially pathogenic organisms; thus the cell wall must be highly resistant, durable and strong (Scheller & Ulvskov, 2010). In order to achieve this, cell walls are organised in a highly-ordered and tightly-packed intricate structure (Figure 1.4), where the crystalline cellulose is coated with hemicelluloses via hydrogen

bonds (maintaining cell wall flexibility by preventing microfibrils adhering to each other), and both polysaccharides are embedded in lignin (adding strength and rigidity to the cell walls) which is cross-linked to hemicellulose via ferulic acid ester linkages (Viikari *et al.*, 2012).

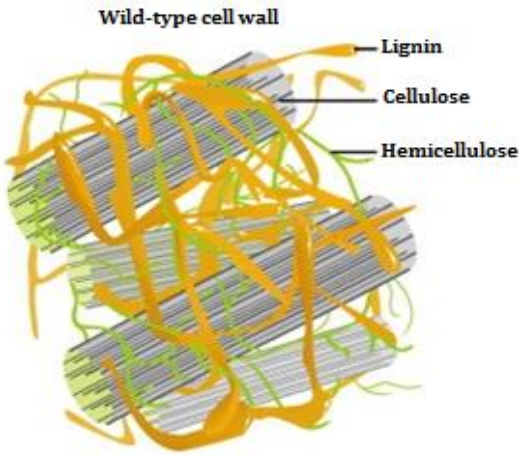


Figure 1.4. Characteristic organization of lignocellulosic constituents: cellulose, hemicellulose and lignin. Image adapted from Vanholme *et al.*, 2010.

The complexity, compactness and recalcitrance of the plant cell wall structure limits the rate and extent of utilisation of biomass materials, representing the most difficult challenge (and economic barrier) to overcome in producing biofuels and industrial chemicals from lignocellulosic feedstocks (Himmel *et al.*, 2007).

1.3.2. Pretreatment

Contrary to first-generation bioethanol production, where the sucrose- and starch-based raw materials are readily hydrolysed into fermentable sugars, the production of lignocellulosic bioethanol requires an additional and indispensable step of pretreatment before hydrolysis to overcome the recalcitrance of plant biomass. Pretreatment steps are required to facilitate rapid and efficient hydrolysis of carbohydrates to fermentable sugars by increasing the surface area of the available cellulose through (i) removing the lignin seal, (ii) solubilising hemicellulose, (iii) disrupting the crystallinity of cellulose, and/or (iv) increasing pore volume (Demain *et al.*, 2005).

There are several pretreatment methods including: (a) physical pretreatment such as grinding and milling, microwave irradiation, pyrolysis, extrusion and freezing; (b) chemical pretreatment using alkali, concentrated or diluted acids, organic solvents,

hydrogen peroxide, ethylene diamine, ozonolysis or ionic liquids; (c) physico-chemical pretreatments like steam explosion, liquid hot water, ammonia fibre explosion (AFEX), wet oxidation and super-critical carbon dioxide explosion; and (d) biological pretreatment using white-, brown- and soft-rot fungi (Chiaramonti *et al.*, 2012; Haghighi Mood *et al.*, 2013; Sun & Cheng, 2002; Viikari *et al.*, 2012).

In each of these methods, the biomass is reduced in size and its physical structure is opened; however, they differ in their advantages and disadvantages with regard to the different lignocellulosic feedstocks used, which need to be considered when trying to find the best pretreatment. In addition, regardless of the pretreatment method(s) used, one drawback is the generation of some by-products such as furan derivatives from sugar degradation (*e.g.* furfural and 5-(hydroxymethyl)furfural (HMF)), weak aliphatic acids (*e.g.* levulinic, acetic and formic acids) and the phenolic compounds from lignin degradation that are considered to be inhibitory to fermentative microorganisms (Haghighi Mood *et al.*, 2013; Viikari *et al.*, 2012).

After the pretreatment step, the lignocellulosic polysaccharides need to be hydrolysed into their constituent sugars, prior to their fermentation to ethanol, which then has to be separated and purified, usually by a process of distillation–rectification–dehydration (Mussatto *et al.*, 2010). This hydrolysis can be achieved chemically using acids or biologically using enzymes. Enzymatic hydrolysis is favoured over acid hydrolysis due to its lower energy consumption, mild operating conditions, high sugar yields, and lower capital and maintenance costs of equipment (Petrova & Ivanova, 2014).

1.3.3. Enzymatic hydrolysis

Lignocellulose represents the most abundant carbohydrate source for microorganisms that have adapted to extract energy and metabolic building blocks from plant biomass (Zeigler, 2014). These plant degraders are found either free in the environment or within the digestive track of xylophagous (wood-feeding) insects (*e.g.* termites and bark and wood-boring beetles) and herbivorous animals. Due to the recalcitrant and complex heterogeneous structure of the plant cell wall, Nature has equipped these microbes with a variety of hydrolytic enzymes, along with different physiological schemes for the degradation of the plant cell wall polysaccharides. In anaerobic bacteria, such as *Clostridia*, multi-enzyme complexes, named cellulosomes, comprise a set of enzymes and substrate-binding modules associated with the cell surface that mediate cell attachment to the insoluble substrate and degrade it to

soluble products that are then absorbed. On the other hand, in the case of aerobic bacteria or filamentous fungi (*e.g. Trichoderma* and *Aspergillus*), a large variety of free extracellular enzymes are secreted to completely degrade the polymers outside the cell (Blumer-Schuette *et al.*, 2014; Shulami *et al.*, 2014).

In both cases, the complete deconstruction of the plant cell wall polysaccharides to their constituent sugars requires a large number of enzymes to liberate the free hexose and pentose sugars, for their subsequent fermentation to bioethanol. The enzymes responsible for the complete degradation of cellulose and hemicellulose (cellulases and hemicellulases, respectively) mainly correspond to glycoside hydrolases (EC 3.2.1.-), which are a widespread group of enzymes that hydrolyse the glucosidic bond between two or more carbohydrate units or between a carbohydrate and a non-carbohydrate moiety (Henrissat *et al.*, 1995). Many of these enzymes are modular proteins with at least two distinct modules: the catalytic module and a carbohydrate-binding module (CBM), the latter of which is thought to have one or more of the following functions: enzyme concentration on the surface of the substrate, substrate targeting/selectivity, and disruption of non-hydrolytic crystalline substrates, all of which properties facilitate and enhance the enzymatic activity of glycosyl hydrolases (Arantes and Saddler 2010).

According to their specific characteristics such as amino acid sequence, mechanism, and structure, these enzymes are currently classified into 133 families, and because the fold of proteins is better conserved than their sequences, some of the families can be grouped in 'clans'. This information is permanently available and updated through the Carbohydrate-Active Enzymes (CAZy) database (Lombard *et al.,* 2014).

1.3.3.1. Cellulases

Cellulases (Figure 1.5) are distinguished from other glycoside hydrolases by their ability to hydrolyse β-1,4-glucosidic bonds between glucosyl residues and include three types of enzymes that cooperatively degrade cellulose (Demain *et al.*, 2005):

- *Endo-1,4-β-D-glucanases* (EC 3.2.1.4), hereafter referred to as endoglucanases, which randomly cleave internal β-1,4-glucosidic bonds at amorphous sites in the cellulose polysaccharide chain, generating oligosaccharides of various lengths and consequently new chain ends. To date, these enzymes are found in the GH families 5, 6, 7, 8, 9, 10, 12, 26, 44, 45, 48, 51, 74, and 124 in the CAZy database.

- Two different types of exoglucanases:
  - *Exo-1,4-β-D-glucanases* (EC 3.2.1.74), which attack the non-reducing end of cellulose to yield cellobiose as the primary product, and liberate D-glucose from β-glucan and cellodextrins. These enzymes can be found in the GH families 1, 3, 5 and 9 in the CAZy database.
  - *Exo-1,4-β-D-glucan cellobiohydrolase* (EC 3.2.1.91), which release cellobiose either from the reducing or non-reducing end of cellulose and liberate D-cellobiose from β-glucan in a processive manner. These enzymes can be found in the GH families 5, 6 and 9 in the CAZy database.

- *β-D-glucosidases* (EC 3.2.1.21), which release D-glucose units from soluble cello-oligosaccharides, cellodextrins and a variety of glycosides. These enzymes are found in the GH families 1, 3, 5, 9, 30 and 116 in the CAZy database and are very important as insufficient β-glucosidase activity causes an accumulation of cellobiose with the consequence that both endoglucanase and exoglucanase activities are severely inhibited (Jeng *et al.*, 2011)

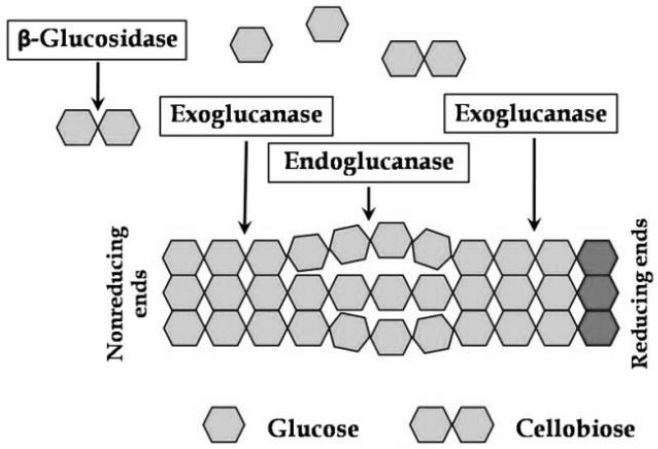


Figure 1.5. Schematic representation of the enzyme system involved in the degradation of cellulose. Image obtained from Ratanakhanokchai *et al.*, 2013.

1.3.3.2. Hemicellulases

Hemicellulases (Figure 1.6) comprise a large variety of cooperatively acting enzymes responsible for the degradation of heterogeneous hemicelluloses (Collins *et al.*, 2005), which also have a synergistic effect in the degradation of cellulose by making this more accessible and vulnerable to attack by cellulases. Within the different types of hemicelluloses, heteroxylans are the most abundant, and the main enzymes involved in its backbone degradation are:

- *Endo-1,4-β-D-xylanases* (EC 3.2.1.8), hereafter referred to as xylanases, which randomly hydrolyse the β-1,4 bond in the xylan backbone, yielding short xylo-oligomers. To date these enzymes are found in the GH families 5, 7, 8, 9, 10, 11, 12, 16, 26, 30, 43, 44, 51 and 62 in the CAZy database.

- *β-D-xylosidases* (EC 3.2.1.37), which are exo-type glycosidases that catalyse the successive removal of xylosyl residues from the non-reducing termini of xylobiose and linear xylo-oligosaccharides. These enzymes are found in the GH families 1, 3, 30, 39, 43, 51, 52, 54, 116 and 120 in the CAZy database.

In addition, the xylan backbone is generally highly substituted with residues of arabinose, glucuronic acid or its 4-O-methyl ether, and acetic, ferulic, and *p*-coumaric acids, all of which can be steric obstacles to the action of xylanases and β-xylosidases and thus limit the hydrolysis of the xylan backbone (Minic *et al.*, 2004). Therefore, for complete hydrolysis to occur, the side chains must be cleaved by several auxiliary debranching hemicellulases, including:

- *α-L-arabinofuranosidases* (EC 3.2.1.55), which are exo-type glycosidases that catalyse the successive removal of arabinose residue from the non-reducing termini of α-1,2-, α-1,3- and α-4,6-linked arabinofuranosyl residues. These enzymes can be found in the GH families 2, 3, 10, 43, 51, 54, and 62 in the CAZy database.

  It should be noted that, due to the spatial similarity between D-xylopyranose and L-arabinofuranose (Appendix IIa,b), the glycosidic bonds and hydroxyl groups can be overlaid leading to bifunctional β-xylosidase/α-arabinofuranosidase enzymes that can be currently found mainly in GH families 3, 43 and 54; these are of great industrial interest due to the presence of synergetic activities in the same active site (Wagschal *et al.*, 2009).

- *α-D-glucuronidases* (EC 3.2.1.139), which catalyse the hydrolysis of α-1,2 glycosidic bonds between 4-O-methyl-D-glucuronic acid and xylan. So far, these enzymes can be found in the GH families 4 and 67.

- *Carbohydrate esterases* are enzymes that catalyse the de-O or de-N-acylation of substituted saccharides.
  - Acetyl xylan esterases (EC 3.1.1.72) hydrolyse the ester linkages between xylose units of xylan and acetic acid and can be found in the CE families 1, 2, 3, 4, 5, 6, 7, 12 and 15 in the CAZy database.

- Feruloyl esterases (EC 3.1.1.73) that hydrolyse the ester linkages between arabinofuranosyl side-chain residues and ferulic acid; these enzymes are found only in the CE family 1.
- *p*-Coumaroyl esterases (EC 3.1.1.-) that hydrolyse the ester linkages between arabinofuranosyl side-chain residues and *p*-coumaric acid.

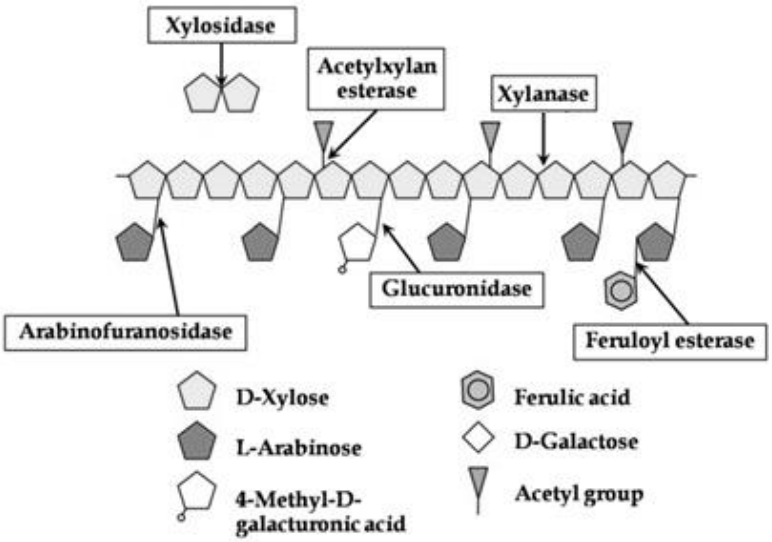


Figure 1.6. Schematic representation of the enzyme system involved in the degradation of heteroxylan. Image obtained from Ratanakhanokchai *et al.*, 2013.

1.3.3.3. Thermophilic carbohydrate degrading enzymes

As discussed above, the efficient saccharification of lignocelluloses into fermentable sugars requires a whole set of biocatalysts, which brings substantial economic implications due to cost of these commercial enzymes. For this reason, it has been increasingly suggested that a mixture of carbohydrate-degrading enzymes derived from plant-degrader thermophilic microorganisms can potentially reduce process costs during the bioconversion of plant cell wall to biofuels. Thermophilic enzymes are both thermostable and thermoactive (allowing them to perform optimally in high-temperature bioprocesses), they are also often tolerant of otherwise harsh industrial conditions for biological systems (including high pressure and protein denaturing solvents) and these unique characteristics allow thermophilic enzymes to be superior in their hydrolytic activity to their mesophilic counterparts (Blumer-Schuette *et al.*, 2014; De Maayer *et al.*, 2014a).

1.3.4. Fermentation

The resulting monosaccharides from the hydrolysis of the plant biomass are then anaerobically fermented by ethanologenic microorganisms (such as the yeast *Saccharomyces cerevisiae*, traditionally used in wine making, baking, and brewing, or the bacterium *Zymomonas mobilis)* to produce high yields of bioethanol. Even though the metabolic pathways and constituent enzymes vary among different microorganisms, in general ethanol is produced from pyruvate, which can be obtained from C6 sugars that follow either the Embden-Meyerhof-Parnas pathway (*e.g. S. cerevisiae)* or the Entner-Doudoroff pathway (*e.g. Z. mobilis*), and from C5 sugars able to enter the Pentose Phosphate pathway to produce the intermediates fructose-6-phosphate and glyceraldehyde-3-phosphate that then continue through glycolysis as shown in Figure 1.7.

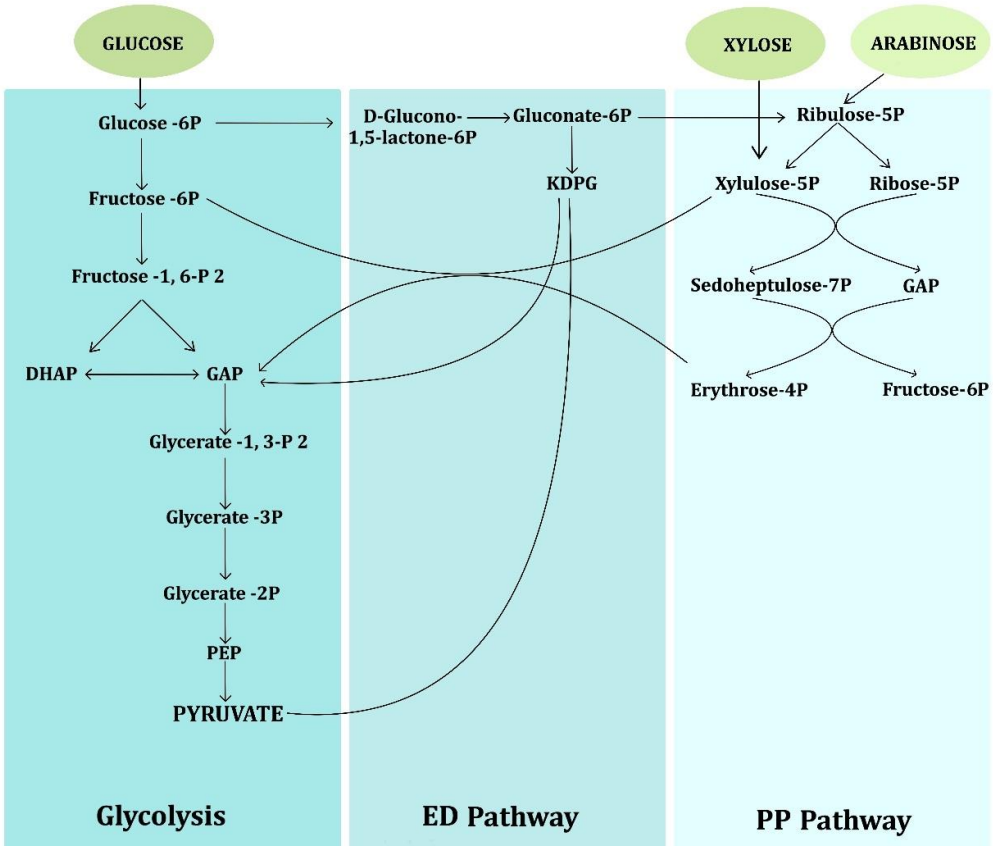


Figure 1.7. Summary of the main central metabolic pathways. (a) Embden-Meyerhof-Parnas. (b) Entner-Doudoroff. (c) Pentose Phosphate. Glucose is phosphorylated to glucose-6-phosphate by a hexokinase and arabinose is converted to ribulose-5-phosphate by an arabinose isomerase and a ribulokinase. Xylose is converted into xylulose-5-phosphate by either the action of xylose reductase, xylitol dehydrogenase and xylulose kinase or by a xylose isomerase and xylitol kinase.

Pyruvate can be converted to acetaldehyde, which is then reduced to ethanol by the enzyme alcohol dehydrogenase, using one of the following two pathways: (1) A one-step pathway that is used by yeasts like *S. cerevisiae* and a few bacteria such as *Z. mobilis*, where pyruvate is non-oxidatively decarboxylated to acetaldehyde, catalysed by pyruvate decarboxylase; or (2) A two-step pathway (more widespread in bacteria) where pyruvate is oxidatively decarboxylated to acetyl-coenzyme A (acetyl-CoA) by the pyruvate dehydrogenase complex (PDHC), by pyruvate ferredoxin oxidoreductase (POR), and/or by pyruvate formate lyase (PFL); acetyl-CoA is then converted to acetaldehyde by a CoA-dependent-acetylating acetaldehyde dehydrogenase (AldDH) (Eram & Ma, 2013).

To date, no single naturally-occurring ethanologenic microorganism has been found that meets the rates, yields and titres for bioprocessing, and the economic targets required for the production of bioethanol and other value-added chemical products from biomass. Well-studied microorganisms like *S. cerevisiae* and *Clostridium thermocellum* are unable to metabolise the pentoses released from hemicelluloses, thus limiting the amount of biomass-derived sugars that can ultimately be reduced to alcohols (Blumer-Schuette *et al.*, 2014). On the other hand, microorganisms that are able to co-metabolise both C6 and C5 sugars, produce minimal amounts of ethanol, making its recovery not economically viable; or have low ethanol tolerance (*e.g. Pichia stipitis*), limiting their growth. For these reasons, extensive efforts have been made by various research groups to isolate novel ethanologenic microorganisms or to introduce heterologous missing pathways into suitable microbial hosts (Chung *et al.*, 2014; Cripps *et al.*, 2009).

1.3.4.1. Thermophilic ethanol producers

As high temperatures are often required for the pretreatment and delignification of plant biomass, there has been increasing interest in the use of thermophilic ethanol-producing microorganisms such as some members of the genera *Clostridium, Caldanaerobacter, Thermoanaerobacter*, *Thermoanaerobacterium* and *Geobacillus* (Blumer-Schuette *et al.*, 2014; Cripps *et al.*, 2009; Sánchez & Cardona, 2008) as they display a number of important advantages in comparison to mesophilic ethanologens. Thermophiles are catabolically versatile, and their high growth temperatures promote high rates of reaction (enhancing the biomass conversion rate) and reduce the risk of biological contamination, which is especially important in processes using municipal or agricultural waste (Blumer-Schuette *et al.*, 2014;

Cripps *et al.*, 2009; Suzuki *et al.*, 2013; Taylor *et al.*, 2008). Moreover, the use of a thermophilic strain can potentially improve the process by helping to reduce the costs associated with the energy input required to cool mesophilic fermentations, while facilitating easy removal of volatile products and recovery by gas stripping, and maintaining anaerobic conditions due to lower oxygen solubility (Cripps *et al.*, 2009; Suzuki *et al.*, 2013; Taylor *et al.*, 2008).

TMO Renewables Ltd was founded in 2002, and in 2008 built the UK's first industrial scale lignocellulosic bioethanol plant (Process Demonstration Unit). Their leading technology was based on a particular strain of the genus *Geobacillus,* which currently comprises nineteen obligate thermophiles. *Geobacillus* spp. are Gram-positive rod-shaped, aerobic or facultatively anaerobic, endo-spore forming microbes (De Maayer *et al.*, 2014a; Zeigler, 2014), and the strain *G. thermoglucosidasius* NCIMB 11955 was isolated from a compost heap. This bacterium was selected by TMO Renewables on account of the wide range of substrates on which it could grow (including polymeric and oligomeric carbohydrates), its ability to ferment both pentose and hexose sugars, and its tolerance to high ethanol concentrations. This wild-type strain grows optimally between 55-65°C, pH 6-8, and crucially produces small amounts of ethanol under fermentative conditions, demonstrating that it possesses the enzymes for bioethanol generation.

TMO Renewables thus developed a genetic toolbox that allowed the molecular engineering of the strain at the metabolic level to produce increased amounts of ethanol. This was possible by disrupting the lactate dehydrogenase gene (*ldh*A) and the pyruvate formate lyase gene (*pfl*B), in addition to up-regulating the expression of the pyruvate dehydrogenase complex by placing the operon under the control of an anaerobically-inducible promoter (the *ldh* promoter from *G. stearothermophilus* NCA1503); these modifications resulted in the patented *G. thermoglucosidasius* triple mutant strain (*ldh*$^{-}$, *pdh*^{up}, *pfl*$^{-}$) identified as TM242. By eliminating the catabolic pathways leading to other end products such as lactate and formate (Figure 1.8), ethanol becomes the major fermentation product, with yields comparable to yeast (*i.e.* greater than 80% of the theoretical yield) (Cripps *et al.,* 2009) suggesting the great potential of this microorganism in the production of bioethanol.

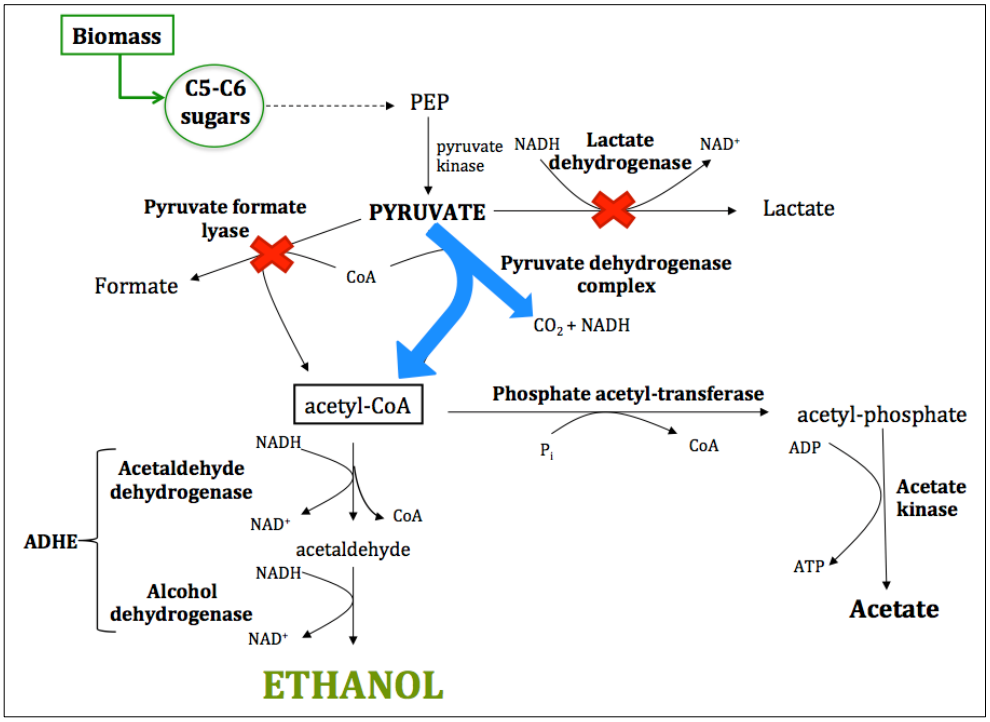


Figure 1.8. Overview of the modified mixed acid fermentation pathways of *G. thermoglucosidasius* TM242. Disruption of *ldh* and *pfl*B genes is shown with red crosses, and the up-regulation of the pyruvate dehydrogenase complex by the green arrow. Image adapted from Cripps *et al.*, 2009.

1.3.5. Consolidated Bioprocessing

Currently, effective pretreatment followed by a high loading of industrial hydrolytic enzymes is needed to achieve an efficient, rapid and complete depolymerisation of the cellulosic and hemicellulosic fraction of biomass. The high cost of the commercial enzymes required constitutes one of the major technical and economical bottlenecks in the overall bioconversion process, hindering the use and commercialisation of lignocellulosic bioethanol (Arantes & Saddler, 2010; Klein-Marcuschamer *et al.*, 2012).

In order to expand the utilisation of lignocellulosic biomass and be cost competitive with grain-derived ethanol, the enzymatic hydrolysis must become more efficient and far less expensive, and for this reason the study of the hydrolytic enzymes required to break down the highly complex plant cell wall has become increasingly important.

A process designed as a possible solution to overcome the economic costs of procuring or producing industrial hydrolytic enzymes is consolidated bioprocessing (CBP), where enzyme production, saccharification, and fermentation of biomass to ethanol is

performed by the same microorganism in one reactor vessel. It has been estimated that such a process could reduce the costs by 40% not considering pretreatment (Lynd *et al.*, 2008; Olson *et al.*, 2012).

Considerable effort has been made to find such CBP microbes, robust in both their enzymatic and metabolic capabilities; however, to date suitable microorganisms with a complete repertoire of all the necessary carbohydrate-degrading enzymes able to ferment both the pentose and hexose sugars have still not been found in Nature. Therefore, it is thought that microorganisms will need to be engineered with the capacity for both general hydrolysis of biomass and simultaneous production of a combustible fuel (Olson *et al.*, 2012)

The major advances in genome and metagenome sequencing achieved in recent years have provided an ever-expanding and biodiverse collection of genes encoding appropriate biocatalysts with the potential of being incorporated into wild-type strains through synthetic biology (Hasunuma *et al.*, 2013). In order to develop a consolidated bioprocessing microorganism, heterologous fermentation pathways or missing genes need to be optimised and introduced into plant biomass degraders through extensive metabolic engineering; or ethanologenic strains need to be engineered with carbohydrate degrading enzymes- and sugar transporter-encoding genes recruited from plant-degraders (Blumer-Schuette *et al.,* 2014). In addition, it has also been suggested that a minimal organism be completely engineered as an alternative to both strategies (Bokinsky *et al.*, 2011).

As described above, *G. thermoglucosidasius* TM242 possesses the capacity to readily ferment both pentose and hexose sugars and it has already been metabolically engineered to optimise the production of ethanol. Furthermore, members of this genus seem to have evolved to tap into the vast global carbon cycle by helping decompose the plant cell wall complex polysaccharides (Zeigler, 2014); TM242 seems to be already equipped with some of the necessary sugar sensors and transporters, as well as some of the hydrolytic enzymes, suggesting the great potential of this microorganism as a platform for CBP.

1.4. Aims of the project

This research was initially supported by TMO Renewables Ltd, a company now partially taken over by ReBio Technologies Ltd who are still interested in biomass conversion to biofuels by the *Geobacillus thermoglucosidasius* microorganism. The project presented in this thesis is part of a collaborative investigation by the Centre for Extremophile Research led by Professor Michael Danson, together with the research group of Professor David Leak, to study and develop *G. thermoglucosidasius* as a consolidated bioprocessing microorganism useful for second-generation bioethanol production.

The aims of my research were:

- To identify the genes encoding enzymes associated with cellulose and hemicellulose hydrolysis in the genome of *G. thermoglucosidasius* TM242.
- To clone and express in *E. coli* some of these relevant glycosidase-encoding genes.
- To characterise the recombinant enzymes obtained in terms of their activity, kinetic parameters, thermostability, thermoactivity and structure.
- To identify in closely-related microorganisms some of the xylanase-encoding genes that are missing from TM242 and to clone and express them in *E. coli* and in *Geobacillus.* Other members of the research group were focussing on improving the cellulolytic ability of this organism using a similar strategy.

The following chapters report Materials and Methods used (Chapter 2), Cloning and expression of TM242 glycosidase-encoding genes and the purification and characterisation of the recombinant enzymes (Chapter 3), X-ray crystal structure solution of the first GH52 β-xylosidase (Chapter 4), Cloning and heterologous expression of closely related *Geobacillu*s xylanases in *E.coli* and TM242 (Chapter 5) and a final overall Discussion and suggestions for on-going research (Chapter 6).

CHAPTER 2: General Materials and Methods

2.1. Microbiology

2.1.1. Bacterial strains

- *Escherichia coli* JM109 (Promega, Southampton, UK). This strain is a non-expression host for general purpose cloning and plasmid propagation as it is endonuclease (*end*A) and recombinase (*rec*A) deficient, which ensures DNA stability and results in high-quality plasmid. These cells are also deficient in β-galactosidase activity due to deletions in both genomic and episomal copies of the *lac*Z gene. The deletion in the episomal (F´ factor) copy of the *lac*Z gene (*lac*ZΔM15) can be complemented by addition of a functional α-peptide encoded by a vector (*e.g.* pGEM®-T Easy).
- Chemically competent *E. coli* BL21(DE3) (Novagen®, Merck Millipore, Watford, UK). This strain is a general purpose expression host as it is deficient in proteases (*Ion* and *omp*T) to favour protein expression. It possesses a lysogen of bacteriophage DE3 and it contains the gene for T7 RNA polymerase under control of the *lac*UV5 promoter.
- *Geobacillus thermoglucosidasius* TM242 (Cripps *et al.*, 2009) supplied by TMO Renewables Ltd. This strain is the *ldh*A⁻*pfl*⁻P_*ldh*/*pdh*^{up} variant of *G. thermoglucosidasius* NCIMB 11955 described in Chapter 1 (section 1.3.1.5).
- *G. thermoglucosidasius* C56-YS93 supplied by Dr David Mead from Lucigen Corporation. This strain has its genome sequence publicly available.
- *Geobacillus* sp. CAMR5420 from the Centre for Applied Microbiological Research, Porton Down (CAMR) Thermophile Culture Collection held at the Centre for Extremophile Research (CER), University of Bath. According to its 16S rRNA analysis, this strain most likely corresponds to the strain *Geobacillus kaustophilus*, and its genome sequence was determined using the Illumina GAIIx platform by the Research Group of Professor Don Cowan at University of Pretoria, South Africa (De Maayer *et al.*, 2014b).
- *Geobacillus* sp. CAMR12739 from the CAMR Thermophile Culture Collection held at the CER. According to its 16S rRNA analysis, this strain most likely corresponds to the strain *Geobacillus stearothermophilus* or *Geobacillus thermoleovorans*, and its genome sequence was determined using the 454 GS-FLX platform by the Research Group of Prof Don Cowan (De Maayer *et al.*, 2014b).

2.1.2. Bacterial growth media and storage

The different media used during this project are summarised below; unless otherwise stated, the media components were supplied by Melford (Ipswich, UK) and salts and buffers were from Sigma-Aldrich Company Ltd (Gillingham, UK). Solid plates of the different liquid media were made by adding 1.5% (w/v) agar prior to sterilisation by autoclaving at 121°C for 15 min (with the exception of Overnight Express™ Instant TB medium (Merck Millipore) that was microwaved according to the manufacturer's instructions). When required, the appropriate selection antibiotic, carbenicillin or kanamycin (at a final concentration of 50 µg/mL for *E. coli* and 12 µg/mL for *Geobacillus*), was added to the media prior to inoculation with a single colony or loop-full of frozen glycerol stock.

2.1.2.1. E. coli

- Lysogeny Broth (LB) medium contains 1% (w/v) NaCl, 1% (w/v) tryptone and 0.5% (w/v) yeast extract.
- Terrific Broth (TB) is a nutritionally-rich medium that increases yields in plasmid-bearing *E. coli,* as it helps the recombinant strains to maintain an extended growth phase. It contains 1.2% (w/v) tryptone, 2.4% (w/v) yeast extract, 72 mM K_2HPO_4, 17 mM KH_2PO_4, and 0.4% (v/v) glycerol.
- Overnight Express™ Instant TB (Merck Millipore) is an autoinduction culture medium for high-level protein production in IPTG-inducible bacterial expression systems (*e.g.* pET). It was prepared with 6% (w/v) instant medium supplemented with 1% (v/v) glycerol.
- Yeast Extract Nutrient Broth (YENB) medium contains 0.75% (w/v) yeast extract and 0.8% (w/v) nutrient broth (BD, Oxford, UK), adjusted to pH 7.5 with NaOH.
- Super Optimal with Catabolite repression (SOC) medium is a rich medium used primarily for the recovery step of *E. coli* after transformations, as it maximises the transformation efficiency of competent cells. It contains 2% (w/v) tryptone, 0.5% (w/v) yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$ and 20 mM glucose.
- SelenoMethionine Medium Complete (Molecular Dimensions Ltd (MDL), Newmarket, UK) is based on a synthetic M9 minimal medium supplemented with glucose, vitamins and amino acids with the exception of L-methionine. It was prepared with 2.16% (w/v) of SelenoMet™ medium base supplemented with 0.51% (w/v) SelenoMet nutrient mix and 0.4% (v/v) 250x SelenoMethionine.

2.1.2.2. Geobacillus sp.

- Soy Peptone Yeast Extract (2SPYNG) Medium: contains 1.6% (w/v) soy peptone (BD), 1% (w/v) yeast extract (Oxoid Ltd, Basingstoke, UK) and 0.5% (w/v) NaCl, adjusted to pH 7.0 with 5 M KOH.
- Modified Urea Salts Medium (USM) contains 0.1% (w/v) yeast extract (Oxoid), 2 mM citric acid·H_2O, 20 mM K_2SO_4, 25mM NaH_2PO_4, 1.25 mM $MgSO_4 \cdot 7H_2O$, 2 mg/L $CaCl_2 \cdot 2H_2O$, 0.4 mg/L Na_2MoO_4, 50 mM urea, 0.012 mM biotin and 5 ml/L of trace elements solution, adjusted to pH 7.0 with 10 M NaOH. The trace elements solutions consisted of 60mM H_2SO_4, 1.44 g/L $ZnSO_4 \cdot 7H_2O$, 1.69 g/L $MnSO_4 \cdot 7H_2O$, 5.56 g/L $FeSO_4 \cdot 7H_2O$, 0.56 g/L $CoSO_4 \cdot 6H_2O$, 0.25 g/L $CuSO_4 \cdot 6H_2O$ 0.08 g/L H_3BO_3 and 0.89 g/L $NiSO_4 \cdot 6H_2O$, dissolved in MilliQ water.

2.1.2.3. Strain storage

To preserve bacteria harbouring expression vectors with verified target gene insert sequences, glycerol stocks were prepared and stored at -80°C in cryogenic vials. Aliquots of 1 mL of overnight cultures, in the required growth medium, were supplemented with sterile glycerol to a final concentration of 20% (v/v). These stocks were mixed thoroughly and snap-frozen using liquid nitrogen prior to immediate transfer to storage.

2.1.3. Preparation of electrocompetent cells

2.1.3.1. E. coli JM109 cells

To prepare electrocompetent *E. coli* JM109 cells, a loop-full of frozen glycerol stock was used to inoculate 10 mL YENB medium. Cells were grown overnight at 37°C in a shaking incubator at 220 rpm. Then 5 mL of this starter culture were used to inoculate 500 mL YENB medium in a 2 L baffled conical flask and incubated at 37°C with shaking at 220 rpm until mid-exponential phase (OD_{600}=0.8-1). The cells were harvested by centrifugation at 6,000g for 10 min at 4°C, and then a series of washing steps were performed on ice using chilled equipment and solutions. The cell pellet was gently resuspended in 50 mL sterile MilliQ water and centrifuged again; this wash step was repeated and then the cells were washed and centrifuged twice using 30 mL of 10% (v/v) sterile glycerol. The final cell pellet was resuspended in 1 mL 10% (v/v) glycerol and 40 µL aliquots of the cell solution were snap-frozen in dry ice and stored at -80°C.

2.1.3.2. Geobacillus thermoglucosidasius TM242 cells

To prepare electrocompetent *G. thermoglucosidasius* TM242 cells, a loop-full of frozen glycerol stock was used to inoculate 5 mL pre-warmed (60°C) 2SPYNG medium. Cells were grown overnight at 60°C in a shaking incubator at 250 rpm. Then 1 mL of this starter culture was used to inoculate 50 mL pre-warmed (60°C) 2SPYNG medium in a 250 mL baffled conical flask and incubated at 60°C with shaking at 250 rpm until early-exponential phase (OD_{600}=1.4-2). The cells were harvested by centrifugation at 4,000g for 20 min at 4°C, and then a series of washing steps were performed on ice using chilled equipment and solutions. The cell pellet was gently resuspended in 50 mL filter-sterilised electroporation medium (0.5 M sorbitol, 0.5 M mannitol, 10% (v/v) glycerol) and centrifuged again; this wash step was repeated three times. The final cell pellet was resuspended in 350 µL electroporation medium, and 60 µL aliquots of the cell solution were snap-frozen in dry ice and stored at -80°C.

2.2. Molecular biology

2.2.1. Polymerase chain reaction (PCR)

2.2.1.1. Gene amplification

To amplify the genes of interest, PCR reactions were performed using Phusion® High-Fidelity DNA Polymerase (Thermo Fisher Scientific, Cramlington, UK) following the conditions recommended by the supplier. A standard PCR reaction (25-50 µL) contained 1x HF Phusion buffer, 200 µM dNTPs (dATP, dCTP, dGTP, dTTP) (Promega), 0.25 µM of each forward and reverse specific primers (detailed in 3.2.3 and 5.2.3), an appropriate amount of DNA template (1 pg–10 ng for plasmid DNA and 50–250 ng for genomic DNA), and 0.5 U PhusionHF DNA polymerase. The reactions were carried out in thin-walled PCR tubes placed into an Eppendorf Mastercycler® gradient PCR thermocycler (Hamburg, Germany), following the programme described in Table 2.1.

Table 2.1. Temperature and time used for each PCR amplification step. Temperature X depends on the annealing temperature of the specific pair of oligonucleotides used. The elongation time (Y) is dependent on the size of the target gene to be amplified (30 s per 1 kb).

	Step	Temperature	Time
	Initial Denaturation	98 °C	30 s
35 cycles	Denaturation	98 °C	10 s
	Annealing	X°C	45 s
	Elongation	72°C	Y min
	Final Elongation	72°C	10 min

2.2.1.2. Colony PCR

To identify positive clones from a large number of transformants, colony PCR reactions were performed using Taq DNA polymerase (GeneSys Ltd, Surrey, UK). A standard PCR reaction (25 µL) contained 0.25 µM of each forward and reverse specific oligonucleotide (the same used previously to amplify the gene of interest), 1x Taq Master Mix (GeneSys Ltd) (which contains appropriate amounts of the Taq DNA polymerase, Taq buffer, dNTPs and loading dye) and DNA template. The genomic DNA was introduced to the PCR mixture reaction by the addition of a small amount of each selected colony, in that the cells were lysed and DNA template released during the initial denaturation step. The reactions were carried out in thin-walled PCR tubes placed into an Eppendorf Mastercycler® gradient PCR thermocycler, following the programme described in Table 2.2.

Table 2.2. Temperature and time used for each colony PCR step. Temperature X depends on the annealing temperature of the pair of oligonucleotides used. The elongation time (Y) is dependent on the size of the target gene to be amplified (1 min per 1 kb).

	Step	Temperature	Time
	Initial Denaturation	95°C	30 s
40 cycles	Denaturation	95°C	10 s
	Annealing	X°C	45 s
	Elongation	72°C	Y min
	Final Elongation	72°C	10 min

2.2.2. Agarose gel electrophoresis

To visualise DNA in the form of PCR products or restriction enzyme digests, agarose gels were prepared by dissolving 0.8-1.2% (w/v) agarose in TAE buffer (40 mM Tris-acetate pH 8.0, 1 mM EDTA) by heating in a microwave until boiling. The solution was then allowed to cool and 0.5 µg/ml ethidium bromide (Sigma-Aldrich) was added. This solution was poured into a gel cassette, a comb was positioned and the gel allowed to polymerise before being placed in a gel tank and covered with TAE buffer. DNA samples were prepared in 6x DNA loading buffer (50% (v/v) glycerol, 50 mM EDTA pH 8.0, 0.05% (w/v) bromophenol blue) and the required volume (5-60 µL) was loaded onto the gel. To determine the approximate size of the DNA, 6 µL of 1 kb DNA ladder (Promega) was loaded with the samples. The electrophoresis was performed at a constant 80-100 V and monitored by following the dye front. The gels were run until the DNA bands were correctly separated. The DNA intercalated with ethidium bromide was visualised using an UV transilluminator.

2.2.3. Gel purification

To purify DNA from agarose gels, the DNA bands were carefully excised under UV light and purified using the Wizard® SV Gel and PCR DNA Clean-Up System (Promega) following the manufacturer's instructions. An elution volume of 35 µL MilliQ water was used and the DNA was stored at -20°C.

2.2.4. pGEM®-T Easy vector cloning

All the PCR-amplified genes used during this project were first cloned into the pGEM®-T Easy vector (Promega) following the protocol described below. The clones confirmed to harbour pGEM®-T Easy with the target gene inserts were then subcloned into the bacterial expression vectors: pET-28a(+) (Novagen®) and pUCG4.8 (kindly donated by PhD candidate Ali Hussein) for heterologous expression in *E. coli* and *Geobacillus,* respectively.

The cloning vector pGEM®-T Easy was selected as it has a single 3´-terminal thymidine at both ends, which greatly improves the efficiency of ligation of PCR products by preventing recircularization of the vector and providing a compatible overhang for A-tailed PCR products. It is a high-copy-number vector that allows high yield amplification of the target genes. It also contains T7 and SP6 RNA polymerase promoters flanking a multiple cloning site region within the α-peptide coding region of the enzyme β-galactosidase (Appendix I.a).

2.2.4.1. A-tailing of blunt-ended PCR products

The PCR products generated by Phusion® HF are blunt-ended; therefore, they require the addition of a 3' adenine before being efficiently ligated into pGEM®-T Easy vector. Gel-purified PCR fragments (as described in 2.2.3) were A-tailed using Taq DNA polymerase (GeneSys Ltd) following the instructions detailed in the pGEM®-T Easy Vector System Technical Manual. A standard A-tailing reaction (10 µL) contained 1x Taq polymerase buffer, 7 µL DNA, 0.2 mM dATP and 0.5 U Taq polymerase. The reaction mixture was incubated at 70°C for 30 min and the A-tailed DNA was purified using the Promega Wizard® SV Gel and PCR DNA Clean-Up System following the manufacturer's instructions.

2.2.4.2. TA ligation

The A-tailed PCR products (amplified target genes) were ligated into pGEM®-T Easy vector using T4 DNA ligase (Promega). A standard ligation reaction (10 µL) contained 1x T4 DNA ligase buffer, 7.5 µL purified DNA fragment, 1.5 µL vector

pGEM®-T Easy vector and 1 U T4 DNA ligase. The reaction mixture was incubated overnight at 16°C, along with a control ligation containing vector with no insert DNA.

2.2.4.3. Ethanol precipitation

To remove salts and concentrate the plasmids before being transformed into the electrocompetent cells, the 10 µL ligation reaction mixture was precipitated by adding 1 µL of dextran blue (0.15 g dextran blue, 0.5 mL 0.5 M EDTA pH 8.0, 9.5 mL MilliQ water), 0.1 M sodium acetate pH 5.2, and 68 µL of ice-cold absolute ethanol. After mixing, the reactions were incubated at -20°C for 30 min prior to centrifugation at 14,000g for 20 min at 4°C to pellet the DNA. The supernatant was removed and the precipitated DNA was washed with 100 µL of ice-cold 70% (v/v) ethanol and the sample was centrifuged again; this washing step was repeated twice. The precipitated DNA was then left to dry before it was resuspended in 5 µL MilliQ water.

2.2.4.4. Transformation of E. coli JM109 by electroporation

To transform electrocompetent JM109 cells (prepared as described in 2.1.3.1) with the ligation reaction (plasmid with the target gene insert), 0.5 µL of ethanol-precipitated DNA was added to 40 µL of electrocompetent cells and incubated on ice for 5 min. These cells were then transferred into chilled 10 mm path length Gene Pulser® electroporation cuvettes (Bio-Rad Laboratories Ltd, Hertfordshire, UK) and placed in the Micropulser electroporator (Bio-Rad) using the Ec1 setting for *E.coli* (5 ms electrical pulse of approximately 1.8 kV). The transformed cells were resuspended in 1 mL of SOC medium and incubated with shaking at 220 rpm at 37°C for 1 h.

2.2.4.5. Blue/white screening

E. coli JM109 and pGEM®-T Easy vector are compatible with blue/white screening, which takes advantage of intracistronic α-complementation to regenerate β-galactosidase activity. The functional *lac*Z gene product, β-galactosidase, is produced when the *lac*Z coding information missing on the F´ episome in JM109 cells is provided by the plasmid pGEM®-T Easy. This activity was detected by plating 100-200 µL of bacteria, transformed with pGEM®-T Easy vector ligated with the target genes, on LB agar plates supplemented with carbenicillin, previously spread with 100 µL of 100 mM isopropyl β-D-1-thiogalactopyranoside (IPTG; an inducer of the *lac* promoter) (Melford) and 20 µL of 50 mg/mL 5-bromo-4-chloro-3-indolyl β-D-galactopyranoside (X-GAL; a dye that produces a blue color when hydrolysed by β-galactosidase)(Melford) and incubating them overnight at 37°C. If the open reading

frame of the α-peptide is disrupted by insertion of the target gene into the multiple cloning site of the pGEM®-T Easy vector, α-complementation does not occur, and the bacterial colonies remain white, allowing easy identification of transformants carrying the successfully ligated pGEM®-T Easy-target gene constructs.

2.2.5. Plasmid DNA preparation

To extract and purify plasmid DNA, cells harbouring a vector ligated with the target gene (single colonies or a loop-full of frozen glycerol stocks) were inoculated into 10 mL of LB medium supplemented with carbenicillin. Cells were grown overnight at 37°C in a shaking incubator set to 220 rpm. Then the cells were harvested by centrifugation at 6,000g for 10 min at room temperature. Plasmids were subsequently purified using the Promega Wizard® plus SV minipreps DNA purification system, according to the manufacturer's instructions. An elution volume of 50 µL MilliQ water was used and the DNA was stored at -20°C.

2.2.6. DNA sequencing

To corroborate the successful cloning of the target gene sequences into the multiple cloning sites regions of the different vectors used, plasmid DNA was sent for sequencing to Source BioScience LifeSciences (Oxford, UK) or Eurofins MWG Operon (Ebersberg, Germany). The samples were sequenced, in the forward and reverse directions, using the commercially-available oligonucleotides required for each vector (*i.e.* inserts in pGEM®-T easy were sequenced using T7F and SP6 primers while T7F and T7R primers were used in the case of pET-28a(+), and M13F and M13R primers for pUCG4.8).

2.2.7. Expression vectors (pET-28a(+) and pUCG4.8) subcloning

2.2.7.1. Restriction enzyme digests

To linearise the pET-28a(+) and pUCG4.8 expression vectors and generate cohesive ends in the target genes cloned in pGEM®-T Easy, different restriction enzymes were used (detailed in 3.2.5 and 5.2.5). All the enzymes and buffers were supplied by New England Biolabs (NEB) (Hitchin, UK). A standard restriction enzyme digestion reaction (20 µL) contained 2 µL of the recommended buffer, 5-10 µL of purified plasmid DNA, 0.25 U of restriction enzyme(s), and Bovine Serum Albumin (BSA) at a concentration of 100 µg/mL when required. The reaction mixture was incubated for 1-3 h at 37°C or 65°C (depending on each enzyme optimum temperature indicated by the manufacturer).

In the case of pET-28a(+) and pUCG4.8 vectors, the restriction enzymes were inactivated according to the manufacturer's instructions after digestion prior to immediate addition of 0.5 U Shrimp Alkaline Phosphatase (SAP) (Roche Diagnostics Ltd, Sussex, UK) and incubation of the samples at 37°C for another hour. This procedure dephosphorylated the 5' phosphate of the digested DNA to prevent recircularization of the vectors in the absence of an insert.

After the restriction enzyme digests (or SAP-treatment), the samples were loaded onto an agarose gel as described in 2.2.2 and the correct bands were purified following the procedure given in 2.2.3.

2.2.7.2. DNA plasmid ligation

The purified gel bands containing the restriction-digested target genes were ligated into the corresponding SAP-treated expression vectors using T4 DNA ligase (Promega). A standard ligation reaction (10 μL) contained 1x T4 DNA ligase buffer, 0.5-1.5 U T4 DNA ligase and a range of different molar ratios of insert to vector. To determine the amount of plasmid and insert to use, the DNA concentration was estimated using a NanoVue Plus spectrophotometer (GE Healthcare, Chalfont St Giles, UK). The ligation reaction mixtures were incubated overnight at 16°C, after which the DNA was ethanol-precipitated as described in 2.2.4.3 and transformed into *E. coli* JM109 electrocompetent cells as described in 2.2.4.4.

2.2.7.3. Positive clone selection

E. coli JM109 cells, electroporated with the expression vectors ligated with the target genes, were spread onto LB agar plates supplemented with the required selection antibiotic, and incubated overnight at 37°C. To select positive clones, individual colonies were picked using a sterile toothpick, and gridded onto another selective LB agar plate that was also incubated at 37°C. Then these selected colonies were used as template DNA for colony PCR screening performed as described in 2.2.1.2. The PCR products were then loaded onto an agarose gel as described in 2.2.2 and the positive clones were miniprepped following 2.2.5 prior to DNA sequencing as described in 2.2.6.

2.2.7.4. Transformation of the expression hosts

The purified expression vectors with verified target gene insert sequences were transformed into their respective bacterial host cells following the protocols described below.

2.2.7.4.1. Transformation of *E. coli* BL21(DE3) by heat shock

To transform chemically-competent *E. coli* BL21(DE3) cells with purified expression vector pET-28a(+) with the target gene insert, 1 µL of plasmid preparation was added to 20 µL of commercial chemically-competent cells and incubated on ice for 20 min. The samples were then heat shocked at 42°C for 45 s in a water-bath set. The transformed cells were then immediately cooled on ice for at least 3 min and then resuspended in 1 mL of SOC medium and incubated with shaking at 220 rpm at 37°C for 1 h. Then, 100-200 µL of cells were spread onto LB agar plates supplemented with kanamycin at a final concentration of 50 µg/mL and incubated overnight at 37°C to allow colony growth. Finally, a single colony was used to inoculate 10 mL LB medium supplemented with kanamycin; cells were grown overnight at 37°C in a shaking incubator at 220 rpm and the culture used to prepare glycerol stocks as described in 2.1.2.3.

2.2.7.4.2. Transformation of *Geobacillus* by electroporation

To transform electrocompetent *G. thermoglucosidasius* TM242 cells (prepared as described in 2.1.3.2) with the purified expression vector pUCG4.8 with the target gene insert, 2 µL of plasmid preparation were added to 60 µL of electrocompetent cells and incubated on ice for 5 min. These cells were then transferred into chilled 10 mm path-length Gene Pulser® electroporation cuvettes, which were placed in the Micropulser electroporator and subjected to a 5 ms electrical pulse of approximately 2.5 kV (10 µF capacitance and 600 Ω resistance). The transformed cells were resuspended in 1 mL of pre-warmed (60°C) 2SPYNG medium and incubated with shaking at 250 rpm at 60°C for 2 h. Then, 200 µL of cells were spread onto 2SPYNG agar plates supplemented with kanamycin at a final concentration of 12 µg/mL and incubated overnight at 60°C to allow colony growth. Finally, a single colony was used to inoculate 5 mL pre-warmed (60°C) 2SPYNG medium supplemented with kanamycin; cells were grown overnight at 60°C in a shaking incubator at 250 rpm and the culture used to prepare glycerol stocks as described in 2.1.2.3.

2.3. Protein expression and purification

2.3.1. E. coli recombinant protein expression

As detailed in Chapter 3 (Section 3.1.1), the only promoter known to direct transcription of the target gene cloned in the pET-28a(+) plasmid by the host T7 RNA polymerase gene, is the *lac*UV5 promoter, which is inducible by IPTG or by growing the

cells in an autoinducing medium (*e.g.* Overnight Express™ Instant TB) where its components are metabolised differentially to promote automatic induction of protein expression from *lac*-based promoters. These two methods were tested during this project in order to optimise the recombinant expression of glycoside hydrolases and are described below.

2.3.1.1. By induction with IPTG

To express the recombinant proteins by induction with IPTG, a loop-full of frozen glycerol stock of *E. coli* BL21(DE3) harbouring the pET-28a(+)-target gene plasmid, was used to inoculate 10 mL of LB medium supplemented with kanamycin. Cells were grown overnight at 37°C in a shaking incubator at 220 rpm. Then 5 mL of this starter culture were used to inoculate 500 mL of sterile TB medium in a 2 L baffled conical flask, and incubated at 37°C with shaking at 220 rpm until an OD_{600} of 1 was reached, at which point protein expression was induced for 3.5-5 h by the addition of IPTG to a final concentration of 0.3-0.5 mM. Finally, the cells were harvested by centrifugation at 6,000g for 15 min at 4°C, the supernatant was discarded and the pelleted cells were stored at -20°C.

2.3.1.2. By auto-induction

To express the recombinant proteins by auto-induction, a loop-full of frozen glycerol stock of *E. coli* BL21(DE3) harbouring the pET-28a(+)-target gene plasmid, was used to inoculate 10 mL of LB medium supplemented with kanamycin. Cells were grown overnight at 37°C in a shaking incubator at 220 rpm. Then 5 mL of this starter culture were used to inoculate 500 mL of sterile Overnight Express™ Instant TB medium in a 2 L baffled conical flask, and incubated at 37°C with shaking at 220 rpm for 20 h. Finally, the cells were harvested by centrifugation at 6,000g for 15 min at 4°C, the supernatant was discarded and the pelleted cells were stored at -20°C.

2.3.2. Heterologous protein expression in G. thermoglucosidasius TM242

As detailed in Chapter 5 (Section 5.1.1), pUCG4.8 vector has a synthetic, constitutive PUp2n38 (uracil phosphoribosyltransferase) promoter that offers moderate expression without the need for induction.

To express the heterologous proteins, a loop-full of frozen glycerol stock of *G. thermoglucosidasius* TM242 harbouring the pUCG4.8-target gene plasmid, was used to inoculate 5 mL of pre-warmed (60°C) 2SPYNG medium supplemented with kanamycin. Cells were grown overnight at 60°C in a shaking incubator at 220 rpm.

Then 1 mL of this starter culture was used to inoculate 100 mL of pre-warmed (60°C) 2SPYNG medium in a 500 mL baffled conical flask and incubated at 60°C with shaking at 220 rpm for 6 h. Finally, the cells were harvested by centrifugation at 6,000g for 15 min at 4°C, the supernatant was collected, concentrated using an Amicon Ultra 10K centrifugal filter device (Merck Millipore) and stored at 4°C for SDS-PAGE analysis (described in 2.3.4) and enzyme assays of secreted extracellular heterologous protein following 5.2.8; the pelleted cells were stored at -20°C.

2.3.3. Cell extract preparation

To obtain intracellularly expressed soluble recombinant proteins, the pelleted cells were resuspended in 10 mL of cell resuspension buffer (50 mM Tris-HCl, pH 7.4, 500 mM NaCl) containing one EDTA-free complete protease-inhibitor cocktail tablet (Roche). The cells were then lysed by five 15 s bursts (pulses of 20 microns every 30 s) of sonication on ice using a 150 W Ultrasonic Disintegrator (MSE Scientific Instruments, Crawley, UK). The disrupted cells, corresponding to the total cell extract fraction (T), were then centrifuged at 14,000g for 20 min at 4°C. The supernatant, corresponding to the soluble fraction of the cell extract (S), was filtered through 0.45 µm and 0.22 µm Millex® Syringe driven filter units (Merck Millipore) for subsequent purification. The pellet corresponding to the insoluble fraction of the cell extract (I) was then solubilised using 8M Urea (Sigma-Aldrich).

2.3.4. SDS-PAGE

To visualise the recombinantly-expressed proteins, a 10% separating Bis-Tris polyacrylamide gel was prepared by adding 1.67 mL 30% (w/v) acrylamide/bis-acrylamide (19:1 ratio) (Bio-Rad), and 1.91 mL distilled water to a 1.42 mL solution of 3.5x Bis-Tris buffer (1.25 M Bis-Tris pH 6.50–6.8). Gel polymerisation was then initiated by adding 25µl 10% (w/v) ammonium persulphate (APS) (Sigma-Aldrich) and 7 µl tetramethylethylenediamine (TEMED) (Sigma-Aldrich) to catalyse the reaction. The separating gel solution was poured between two glass plates into the casting frame of the Tetra Mini-Protean clamp assembly (Bio-Rad) and overlayed with 70% (v/v) ethanol. The gel was allowed to set before removing the ethanol and rinsing with distilled water.

A stacking gel was prepared by adding 230 µL 30% (w/v) acrylamide/bis-acrylamide and 1.02 mL distilled water to a 500µL solution of 3.5x Bis-Tris buffer prior to initiating gel polymerisation with 20 µl 10% (w/v) APS and 10 µl TEMED. The stacking gel was then added on top of the separating gel, and a comb

was positioned in the gel to create wells. The gel was allowed to polymerise, after which time the comb was carefully removed and the wells were rinsed with distilled water. The complete gel (separating and stacking) was placed in the Tetra Mini Protean apparatus (Bio-Rad) and covered with running buffer (50 mM MOPS, 50 mM Tris, 1 mM EDTA, 0.1% (w/v) SDS, pH 7.7) used for separating proteins larger than 20 kDa.

Samples were prepared by mixing 1-5 µL of protein fraction, 5 µl of the 4x Protein loading buffer (0.2 M Tris-HCl pH 6.8, 2% (w/v) SDS, 40% (v/v) glycerol, 0.1% (v/v) β-mercaptoethanol, 0.8 mg/mL bromophenol blue) and an appropriate volume of distilled water. Broad Range Molecular Weight Standards (Bio-Rad) solution was prepared with 25 µl 4x Protein loading buffer, 5 µl broad range marker and 70 µl distilled water. Both samples and SDS marker solution were denatured by heating at 95°C for 5 min, before loading 10 µl of each onto the gel.

The electrophoresis was performed at 200 V and 400 mA for approximately 45 min or until the dye front had reached the bottom of the gel. Protein bands were visualized by incubating the gel with Coomassie Blue stain solution (45% (v/v) methanol, 10% (v/v) acetic acid and 0.25% (w/v) Coomassie Brilliant Blue R 250 (Sigma-Aldrich)) for 20 min with shaking at 50 rpm. The staining solution was then decanted, the gel was rinsed three times with distilled water and incubated overnight with destaining solution (30% (v/v) methanol, 10% (v/v) acetic acid) with shaking at 50 rpm.

To determine the polypeptide size of the recombinant proteins, the mobility of each protein band was compared with the broad range standard molecular markers. A standard curve was generated by plotting ln (M_r) of the broad range standards (Myosin (200.0 kDa), β-galactosidase (116.2 kDa), Phosphorylase b (97.4 kDa), Serum albumin (66.2 kDa), Ovalbumin (45.0 kDa), Carbonic anhydrase (31.0 kDa), Trypsin inhibitor (21.5 kDa), Lysozyme (14.4 kDa) and Aprotinin (6.5 kDa)) against the relative mobility (Rf), where Rf is equal to the distance migrated by the protein divided by the distance migrated by the dye front.

2.3.5. Protein Purification

2.3.5.1. Immobilised-Metal Affinity Chromatography (IMAC)

To purify recombinant proteins with an engineered poly-histidine affinity tag attached to the N-terminal, IMAC on a Nickel-nitrilotriacetic acid (Ni–NTA) matrix was used.

2.3.5.1.1. Small-scale purification

A plastic polyprep chromatography column (Bio-Rad) was filled with 1 mL chelating cellulose resin (Merck Millipore) and washed with 5 mL of MilliQ water, followed by the addition of 2 mL 0.4 M $NiSO_4$. The column was washed with MilliQ water and equilibrated with His-bind buffer (50 mM Tris-HCl pH 8.0, 300 mM NaCl, 20 mM imidazole (Thermo Fisher)). The small-scale purification was carried out at room temperature and the filtered soluble cell extract (obtained as described in 2.3.3) was added to the column; the loading eluent was collected and loaded onto the column again, with the subsequent first fraction corresponding to the flow through (FT). Following this, any unbound proteins that remained in the column were washed with 5 mL of His-bind buffer (corresponding to 0%). Then the bound proteins (the His-tagged recombinant proteins and other contaminating *E. coli* proteins that contain a high number of histidines) were eluted by sequentially washing the column with 5 mL of increasing concentrations of imidazole (25-500 mM), using different proportions of His-elute buffer (50 mM Tris-HCl pH 8.0, 300 mM NaCl, 1 M imidazole) diluted in His-bind buffer. Finally, undiluted His-elute buffer (corresponding to 100%) was added to clean the column, which was then stored in 5 mL 20% (v/v) ethanol.

After 3 purifications, the nickel was stripped from the column with 2 mL of 0.5 M EDTA pH 8.0; the column was then washed with 5 mL of MilliQ water before being recharged with 2 mL 0.4 M $NiSO_4$. The column was finally washed with 5 mL His-bind buffer and stored with 5 mL 20% (v/v) ethanol.

2.3.5.1.2. Fast protein liquid chromatography (FPLC)

β-Xylosidase-1 was purified by IMAC on an ÄKTA explorer FPLC system (GE Healthcare) at room temperature, using the associated software UNICORN 5.0. All solutions used were first filtered and degassed under vacuum to avoid the presence of bubbles disrupting the system. The soluble cell lysate in cell resuspension buffer (50 mM Tris-HCl pH 7.4, 500 mM NaCl) (obtained as described in 2.3.3) was loaded onto a pre-equilibrated HisTrap HP 5 ml column (GE Healthcare) at a flowrate of 1 ml min^{-1}. The column was washed with 10 column volumes (CV) of loading buffer, followed by a 4 CV wash with the same buffer containing 60 mM imidazole. β-Xylosidase-1 was then eluted with 30% of elution buffer (Tris-HCl pH 8.0, 300 mM NaCl, 500 mM imidazole), and the column was finally washed with a gradient of 60% and 100% of the same buffer. Fractions containing β-xylosidase activity (measured as described in 3.2.8) were pooled and concentrated using an Amicon Ultra 10K centrifugal filter device prior to further purification.

2.3.5.2. Size exclusion chromatography (SEC)

To purify β-xylosidase-1 further, SEC was performed on an ÄKTA explorer FPLC system at room temperature, using the associated software UNICORN 5.0. All solutions used were first filtered and degassed under vacuum to avoid the presence of bubbles disrupting the system. A maximum of 500 µL of concentrated protein sample (from the previous IMAC purification described in 2.3.5.1.2) was loaded onto a Superdex 200 10/300 GL column (GE Healthcare) (24 mL bed volume) column that had been previously equilibrated with 50 mM Tris-HCl buffer, pH 8.0, containing 150 mM NaCl, at a flow rate of 0.2 ml min^{-1}. Proteins were eluted in the same buffer according to their size. Fractions containing β-xylosidase activity (measured as described in 3.2.8) were visualised by SDS-PAGE as described in 2.3.4, and those with the highest purity (~95%) and preferably monodisperse, were concentrated to 20 mg/mL and utilised for the crystallisation studies described in Chapter 4.

To determine the native M_r value (and therefore the oligomeric state) of the eluted protein peaks, standards proteins from a Gel Filtration Markers Kit (Sigma-Aldrich) (Carbonic anhydrase (29 kDa), Bovine albumin (66 kDa), Alcohol dehydrogenase (150 kDa), β-amylase (200 kDa), Apoferritin (443 kDa) and Thyroglobulin (669 kDa)) were similarly chromatographed using the same column, elution buffer and flow rate. A standard curve was generated by plotting log (M_r) versus K_{av}; where K_{av}=(Ve-Vo)/(Vc-Vo) and Ve is the elution volume, Vo is the column void volume and Vc is the geometric column volume.

2.3.6. Estimation of protein concentration

2.3.6.1. Total protein concentration of impure samples

Total protein concentration of impure samples was determined by the Bradford method (Bradford, 1976) using PierceBSA (Thermo Fisher Scientific) as a standard (2-10 µg/ml). Protein assay Dye Reagent (Bio-Rad) was diluted 1 in 5 in distilled water and 900 µL were added to 100 µl of sample (either protein fraction or standards), mixed and incubated at room temperature for 20 min to develop the reaction. Absorbance was read at 595 nm on a Varian Cary 50 Bio UV/visible light spectrophotometer (Varian, Inc, USA) and the total protein concentration of each sample was determined from the standard curve created with the results from the BSA standards.

2.3.6.2. Concentration of pure proteins

The concentration of purified proteins was determined using a NanoVue Plus spectrophotometer (GE Healthcare). The option Protein A_{280} was selected and, with the known molar absorption coefficient (ε_M) and relative molecular mass (M_r) of the purified protein, the concentration was read directly. The values of M_r and ε_M were calculated from the amino acid sequence of each recombinant protein using the ProtParam tool from the Expasy, Swiss Institute of Bioinformatics website (Gasteiger *et al.*, 2003).

CHAPTER 3: Identification, cloning and expression of glycosidase-encoding genes from *Geobacillus thermoglucosidasius* TM242 and characterisation of the recombinant enzymes.

3.1. INTRODUCTION

As described in Chapter 1, biomass is abundant, of low cost and renewable; however, plant cells walls have evolved to be recalcitrant or extremely resistant to degradation. In order to break down the carbohydrate polymers from lignocellulosic feedstock, cellulose and hemicellulose, into their simple C6 and C5 sugars respectively, specific carbohydrate-degrading enzymes (glycoside hydrolases) can be used. The enzymes associated with cellulose and hemicellulose hydrolysis are diverse and act in synergy, which means that several different types of enzymes are required during the enzymatic hydrolysis of biomass: cellulases (endoglucanases, exoglucanases, β-glucosidases) and hemicellulases (xylanases and β-xylosidases, along with auxiliary debranching enzymes: α-arabinofuranosidases, α-glucuronidases, acetyl xylan esterases, and ferulic acid esterases among others).

This chapter describes the search for *G. thermoglucosidasius* TM242's glycoside hydrolases involved with cellulose and hemicellulose hydrolysis, as well as the gene cloning, protein expression and characterisation of some of the recombinant enzymes. The genes selected were those encoding: two β-xylosidases (E.C 3.2.1.37), which are exo-type glycosidases that hydrolyse xylose monomers from the non-reducing end of xylo-oligosaccharides and xylobiose into single xylose units; one α-arabinofuranosidase (EC 3.2.1.55), which is an exo-type glycosidase that catalyses the successive removal of arabinose residue from the non-reducing end of arabinose-containing polysaccharides; one β-glucosidase (EC 3.2.1.21), which hydrolyses glycosidic bonds to release non-reducing terminal glucosyl residues from glycosides and oligosaccharides; a predicted glycoside hydrolase with unknown function, and a β-N-acetylhexosaminidase (EC 3.2.1.52) that catalyses the hydrolysis of terminal non-reducing N-acetyl-D-hexosamine residues in N-acetyl-β-D-hexosaminides and is involved in peptidoglycan turnover and cell wall recycling in bacteria (Litzinger *et al.*, 2010).

3.1.1. Cloning and recombinant expression

For the cloning of TM242's glycosidase-encoding genes, the expression vector of choice was pET-28a(+) (Appendix I.b) due to some of its unique features: it contains

a T7*lac* promoter (a *lac* operator sequence just downstream of the T7 promoter), a kanamycin resistance gene, and the possibility to add an N- or C- terminal six-histidine tag along with a thrombin site to remove the tag if necessary. The chosen host was *E. coli* BL21(DE3), because it is a strain deficient in *lon* protease and the *ompT* outer membrane protease that can degrade proteins during purification (Grodberg & Dunn, 1988); it also possesses a lysogen of bacteriophage DE3, a lambda derivative that has the immunity region of phage 21 and carries a DNA fragment containing the *lac*I gene, the *lac*UV5 promoter, and the gene for T7 RNA polymerase on its chromosomal DNA (Studier & Moffatt, 1986). As this fragment has been inserted into the *int* gene, DE3 is prevented from integrating into or excising from the chromosome without a helper phage. Once a DE3 lysogen is formed, the only promoter known to direct transcription of the T7 RNA polymerase gene is the *lac*UV5 promoter, which is inducible by isopropyl β-D-1-thiogalactopyranoside (IPTG) (Novagen, 2005).

The *E. coli* cells harbouring the pET-28a(+) plasmid carrying each glycosidase-encoding gene must then be grown in a suitable rich medium; this could be non- inducing, thereby requiring the addition of IPTG in order to induce the T7 RNA polymerase that in turn transcribes the target DNA in the plasmid, or an auto-inducing medium that promotes growth to high density and automatically induces protein expression from *lac* promoters, without the need to monitor cell density and without conventional induction with IPTG (Grabski *et al.*, 2005; Studier, 2005). In order to maximise the yield of recombinant proteins, the optimal conditions for bacterial growth and protein expression must be determined empirically; general techniques for optimising heterologous protein overproduction in *E. coli* include the appropriate selection of host strain and plasmid (copy number, selection antibiotic, promoter), growth conditions (volume and composition of culture media, oxygen levels, temperature) and induction (density of cells prior to the induction, concentration of inductor, length of the induction) among others (Donovan *et al.*, 1996; Weickert *et al.*, 1996).

3.1.2. Protein purification and characterisation

After protein expression, the cells are harvested and lysed to obtain a soluble cell extract, from which the recombinant protein with the histidine tag can be purified through Immobilised-Metal Affinity Chromatography (IMAC). The Nickel-nitrilotriacetic acid (Ni–NTA) matrix, which selectively binds the poly-histidine affinity tag attached to the N- or C-terminus, is commonly used (Gaberc-Porekar & Menart, 2001).

During the IMAC purification process, the sample is loaded onto the column under conditions that favour specific binding, particularly a buffer of a pH at which imidazole nitrogens in histidyl residues are in the non-protonated form and of a relatively high ionic strength to reduce non-specific electrostatic interactions; clearly the buffer itself should not coordinatively bind to the chelated metal ion. Elution of the target protein can be achieved by protonation or ligand exchange with imidazole, which often provides a single step isolation of His-tagged recombinant proteins at very high purity (Gaberc-Porekar & Menart, 2001). As with other forms of affinity chromatography, IMAC is used in cases where rapid purification and substantial purity of the product are necessary. However, compared to other affinity separation technologies, it cannot be classified as highly specific, but only moderately, so in cases where higher purity is required (*e.g.* for protein crystallisation), one or more additional purification steps may be added.

In this chapter, once purified, the TM242 recombinantly-expressed enzymes were assayed with a variety of chromogenic *p*-nitrophenyl linked glycosides (Appendix II). On enzymatic hydrolysis of these, *p*-nitrophenol is liberated, which at alkaline pH develops a yellow colour suitable for the quantitative measurement of enzymatic activity spectrophotometrically at a wavelength of 400-420 nm. The active enzymes were also characterised in terms of thermostability, thermoactivity and kinetic parameters.

3.2. MATERIALS AND METHODS

3.2.1. Identification of glycosidases-encoding genes in G. thermoglucosidasius TM242

The glycosidase-encoding genes were identified by comparing the *G. thermoglucosidasius* TM242 genome with known nucleotide sequences of carbohydrate-degrading enzymes mainly involved with cellulose and hemicellulose hydrolysis. The search was predominantly restricted to *Bacillus* related species and sequences were obtained from 28 different GH (Glycoside Hydrolase) families (1, 3, 5, 6, 7, 8, 9, 10, 11, 12, 16, 18, 19, 26, 30, 39, 43, 44, 45, 48, 51, 52, 54, 61, 62, 74, 116, 120) in the CAZy database (Lombard *et al.*, 2014). The online comparison was done using the private database ERGO (Overbeek *et al.*, 2003) which also provided the gene annotation and predicted open reading frame for each sequence.

3.2.2. Bioinformatic analysis of TM242 glycoside hydrolases

The glycosidase-encoding gene sequences found in the *G. thermoglucosidasius* TM242 genome were analysed with two different bioinformatic tools: SignalP4.0 server (Petersen *et al.*, 2011) which predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms, including Gram-positive bacteria, and the Conserved Domain Database (CDD) (Marchler-Bauer *et al.*, 2013) which identifies the protein's conserved domains (including the presence of Carbohydrate Binding Modules (CBMs)), and therefore its putative function and glycoside hydrolase family classification.

3.2.3. Primer design

Forward and reverse primers were designed to be identical to the desired 5' region of the gene of interest and its 3' end, respectively. The primers also incorporated an appropriate restriction site (not present in the gene of interest) at the desired terminus of the DNA sequence. The length of each sequence (26-32 bases) was determined based on similar melting temperatures (T_m).

The nucleotide sequences of six glycosidase-encoding genes from TM242 were used to design the specific primers shown in Table 3.1. These included the restriction sites for NheI (forward) or XhoI (reverse) at each 5' end, as required to clone the amplified genes into the expression vector pET-28a(+) including an N-terminal His tag. In the case of the β-N-acetylhexosaminidase-encoding gene, the signal peptide

sequence was removed in order to ensure solubility, and therefore a start codon was introduced into its forward primer. The primers were all synthesised by Invitrogen Ltd (Paisley, Scotland).

Table 3.1. Primers used to amplify putative glycosidase-encoding genes from TM242. The restriction sites introduced into each primer are underlined. Extra bases introduced to favour the cleavage are shown in lower case. The start codon introduced is shown in bold.

Specific Gene	Primer Name	Sequence	Length	Tm
β-N-acetylhexosaminidase	GH3-F	5' cta<u>GCTAGC</u>**ATG**CAAGCGCATACAGA 3'	26 b	66°C
	GH3-R	5' c<u>CTCGAG</u>TTAGTAAGTGAGACCATGGC 3'	27 b	65°C
α-arabinofuranosidase	ARBF-F	5' cg<u>GCTAGC</u>ATGAAGAATATGAACACAAAAA 3'	30 b	67°C
	ARBF-R	5' c<u>CTCGAG</u>TTTCTTAGCCAAACGAATCA 3'	27 b	67°C
β–glucosidase	BGLU-F	5' g<u>GCTAGC</u>ATGAAGCAACAAACAAAGCAAG 3'	29 b	70°C
	BGLU-R	5' c<u>CTCGAG</u>TTAAAACCGTTGTTCTCCACAAG 3'	31 b	69°C
β–xylosidase-1	BXYL1-F	5' cg<u>GCTAGC</u>ATGCCAAAAAACATGTTTTTTAAC 3'	32 b	70°C
	BXYL1-R	5' ccg<u>CTCGAG</u>TTACTTTCCCTCTTCCAAC 3'	28 b	70°C
β–xylosidase-2	BXYL2-F	5' cg<u>GCTAGC</u>ATGGTAAGAATAAGAAATCCC 3'	29 b	67°C
	BXYL2-R	5' ccg<u>CTCGAG</u>TCACAATTCTTTATACAAAAAG 3'	31 b	67°C
Predicted glycoside hydrolase	PGH-F	5' g<u>GCTAGC</u>ATGGTAAGAGTCAACCGTTTTCTG 3'	31 b	70°C
	PGH-R	5' ccg<u>CTCGAG</u>TTATCGATAAATTAATTTCGCC 3'	31 b	70°C

3.2.4. Amplification of glycoside hydrolases from TM242

G. thermoglucosidasius TM242 was grown aerobically at 60°C for 16 h in 2SPYNG media and its chromosomal DNA was isolated using standard techniques described in Sambrook & Russell (2001). This genomic DNA was used as template for PCR reactions previously described in 2.2.1.1.

3.2.5. Cloning of glycoside hydrolases from TM242

Following PCR amplification, the TM242 glycosidase-encoding genes were purified, A-tailed and ligated into pGEM®-T Easy vector as described in 2.2.3, 2.2.4.1 and 2.2.4.2 respectively. The ligations were ethanol-precipitated as described in 2.2.4.3 and transformed by electroporation into *E. coli* JM109 electrocompetent cells following 2.2.4.4. The positive clones (harbouring pGEM®-T Easy vector containing glycosidase-encoding genes) were selected as described in 2.2.4.5. The correct plasmids were purified following 2.2.5, digested with NheI and XhoI restriction enzymes and then ligated into the NheI/XhoI sites of the expression vector pET-28a(+) multiple cloning site, following the methods previously described in 2.2.7.1 and 2.2.7.2 respectively. These ligations were then ethanol-precipitated and transformed by electroporation into *E. coli* JM109 electrocompetent cells as before and the positive clones (harbouring pET-28a(+) containing glycosidases-encoding genes) were selected as described in 2.2.7.3. Finally, the correct plasmids were purified as before, sequenced following 2.2.6 and then transformed by heat shock into chemically competent *E. coli* BL21(DE3) cells as described in 2.2.7.4.1. To preserve bacteria harbouring expression vectors with

verified insert sequences, glycerol stocks were prepared as described in 2.1.2.3 and stored at -80°C.

3.2.6. Protein expression

The recombinant TM242 glycoside hydrolases were expressed following the method previously described in 2.3.1.

3.2.7. Protein purification

The recombinant His-tagged TM242 glycosidases were purified by small-scale nickel affinity chromatography as described in 2.3.5.1.1 with the exception of β-xylosidase-1, which, due to the high purity requirements for the crystallisation studies shown in Chapter 4, was purified by nickel-affinity and size exclusion chromatography at 25°C on an ÄKTA Explorer FPLC system following the method previously described in 2.3.5.1.2 and 2.3.5.2.

3.2.8. Continuous spectrophotometric enzyme assays

The enzymatic activity of the recombinant glycosidases was routinely assayed spectrophotometrically on a Cary 50 Bio UV/visible light spectrophotometer (Varian, Inc, USA). Activity-appropriate *p*-nitrophenyl glycosides were used as synthetic chromogenic substrates for each enzyme: *p*-nitrophenyl-α-L-arabinofuranoside (Carbosynth, Compton, Berkshire, England), *p*-nitrophenyl-α-L-arabinopyranoside, *p*-nitrophenyl-β-D-xylopyranoside (Glycosynth, Warrington, England), *p*-nitrophenyl-β-D-glucopyranoside and *p*-nitrophenyl-N-acetyl-β-D-glucosaminide (Sigma–Aldrich) (Appendix II). All assays were conducted in a volume of 1 mL, at 65°C, using preheated McIlvaine's buffer (0.16 M Na_2HPO_4, 0.018 M citric acid, pH 7.0), the specific substrate and chilled enzyme, which initiated the continuous assay. The appearance of *p*-nitrophenol was monitored at 410 nm and one unit of enzyme activity is defined as the release of 1 µmol of *p*-nitrophenol per minute under assay conditions. The molar absorption coefficient (ε) of *p*-nitrophenol determined experimentally at 65°C, pH 7.0, was 8,500 M^{-1} cm^{-1}.

3.2.8.1. Further characterisation of β–xylosidase-1

Catalytic activity with the natural substrates, xylobiose and xylotriose, was assayed by determination of the xylose produced using NAD^+ dependent xylose dehydrogenase as a coupling enzyme from the Megazyme D-Xylose Assay Kit (Megazyme, Bray, Ireland). Spectrophotometric assays (1 mL) were carried out at 45°C

in Megazyme buffer, and the production of NADH was followed continuously at 340 nm (the molar absorption coefficient of NADH is 6,220 M^{-1} cm^{-1}). One unit of β-xylosidase activity is defined as the production of 1 µmol of NADH per minute.

3.2.9. Determination of kinetic parameters

Substrate saturation curves were obtained by plotting initial rates against substrate concentration. Analysis of the enzymatic assay results was carried out using the Enzyme Kinetics module in the SigmaPlot 12 Software (Systat Software, Hounslow, England). The kinetic parameters *V*max and K_M for each substrate were determined by non-linear regression, fitting the data to the Michaelis-Menten equation:

$$\text{Michaelis-Menten Equation:} \qquad v = \frac{V\max\,[S]}{K_M + [S]}$$

Where v is the initial velocity, *V*max is the maximum enzyme velocity, [S] is the substrate concentration and K_M is the Michaelis-Menten constant.

3.2.10. Thermostability and thermoactivity

3.2.10.1. Temperature optima

To determine the optimum temperature of each active recombinant enzyme, the assay buffer (McIlvaine pH 7.0) was preheated at different temperatures (20–100°C) for 10 minutes in quartz cuvettes. The enzyme was assayed as described previously at each different temperature.

3.2.10.2. Thermal inactivation

In order to determine a thermal inactivation profile of each active recombinant protein, thermal inactivation assays were conducted in an Eppendorf Mastercycler® gradient PCR thermocycler using thin walled 0.5 mL Eppendorf tubes. The protein was heated in the PCR machine at various temperatures (60-100°C) with the lid set to 105°C to prevent the protein boiling dry. Sample tubes were removed periodically (5, 10, 15, 30 and 60 min) and cooled on ice. Once fully cooled, the PCR tubes were centrifuged at 14,000g for 2 min. The samples were then assayed as described in 3.2.8 to determine the percentage activity of the heat-treated protein compared to the activity of the control protein.

3.3. RESULTS AND DISCUSSION

3.3.1. Identification of carbohydrate-degrading enzymes in Geobacillus thermoglucosidasius TM242.

The bioinformatic analysis of the *G. thermoglucosidasius* TM242 genome resulted in the identification of ten genes encoding enzymes presumed to be involved with cellulose or hemicellulose hydrolysis. The amino acid sequences were then analysed with SignalP4.0 and the CDD to check for signal peptides and the presence of carbohydrate binding modules. A summary of the results obtained is presented in Table 3.2. It should be noted that there are other glycoside hydrolase gene sequences present in the TM242 genome (*e.g.* neopullulanases, exo-1,4-α-D-glucosidases, and amylo-1,6-α-D-glucosidases) that have not been considered as these enzymes are related to starch hydrolysis, in addition to the auxiliary enzymes such as β-mannosidase, α-mannosidase α–galactosidase that do not play a primary role in the hydrolysis of lignocellulosic feedstock.

Table 3.2. *Geobacillus thermoglucosidasius* TM242 putative glycosidase-encoding genes, their predicted function, CAZy classification, amino acid number, and prediction of signal peptides and carbohydrate binding modules. The sequences selected for cloning and expression are shown in bold.

Subject ID	Protein Annotation	GH Family	aa	Signal Peptide	Carbohydrate Binding Module
RTMO 00038	Spore peptidoglycan hydrolase (N-acetylglucosaminidase) (EC 3.2.1.-)	18	476	No	No
RTMO 00277	**Predicted glycoside hydrolase**	**?**	**356**	**No**	**No**
RTMO 01465	**β-N-acetylhexosaminidase (EC 3.2.1.52)**	**3**	**698**	**Yes**	**Maybe**
RTMO 01733	**α-arabinofuranosidase (EC 3.2.1.55)**	**51**	**505**	**No**	**No**
RTMO 01743	**β-xylosidase (EC 3.2.1.37)**	**52**	**704**	**No**	**No**
RTMO 01744	Predicted integral membrane protein	?	257	No	No
RTMO 01745	**β-xylosidase (EC 3.2.1.37) / α-arabinofuranosidase (EC 3.2.1.55)**	**43**	**545**	**No**	**No**
RTMO 01764	**β-glucosidase (EC 3.2.1.21)**	**1**	**473**	**No**	**No**
RTMO 02650	6-phospho- β-glucosidase (EC 3.2.1.86)	1	478	No	No
RTMO 02911	Spore peptidoglycan hydrolase (N-acetylglucosaminidase) (EC 3.2.1.-)	18	428	No	No
RTMO 03636	6-phospho-β-glucosidase (EC 3.2.1.86)	1	478	No	No

Four TM242 genes, annotated as a β–glucosidase, two β–xylosidases and an α-arabinofuranosidase, correspond to glycoside hydrolases that, as described in Chapter 1 (1.3.1.3), are clearly involved in cellulose and xylan oligosaccharides hydrolysis. However, genes encoding xylanases, endoglucanases or exoglucanases were not found in the genome. Only one of the sequences, the one annotated as β-N-acetylhexosaminidase, presented a signal peptide sequence with a predicted

cleavage site between amino acids 27 and 28, therefore indicating that the other glycoside hydrolases found in TM242 are probably intracellular proteins. No sequences contained carbohydrate binding modules.

Six genes were selected for further study, cloning and recombinant expression in *E. coli:* the β-glucosidase (RTMO01764), two β–xylosidases (RTMO 01745 and RTMO 01743), the α-arabinofuranosidase (RTMO 01733), a predicted glycoside hydrolase (RTMO00277) and the β-N-acetylhexosaminidase (RTMO 01465). Even though its function remains unknown, the predicted glycoside hydrolase (RTMO 00277) was chosen because the CDD search indicated that this protein possessed conserved domains that could belong to the GH families 32, 43, 62 or 68, which include enzymes of interest to bioethanol production (*e.g.* xylanases, β-xylosidases and α-arabinofuranosidases). The β-N-acetylhexosaminidase (RTMO 01465) is not related to cellulose or hemicellulose hydrolysis but was selected as it is the only one with a predicted signal peptide, it belongs to GH3 (which also includes β-glucosidases, β-xylosidases and α-arabinofuranosidases) and the CDD search indicates some interesting domains like β-glucosidase-related glycosidases (BglX), probable β-xylosidase (PLN03080) and β-D-glucoside glucohydrolase (PRK15098), which might suggest this protein could be mis-annotated in the ERGO database.

3.3.2. Amplification and cloning of the TM242 glycosidase-encoding genes

The six gene sequences were correctly amplified using the TM242 genomic DNA as template. Specific PCR products of the approximate expected size were obtained as shown in Figure 3.1. The bands were excised from the gel, A-tailed and cloned into pGEM®-T Easy, and then subcloned into the expression vector pET-28a(+). Sequencing confirmed the cloned gene sequences were correct and the vectors harbouring the glycosidase-encoding genes were then transformed into *E. coli* BL21(DE3).

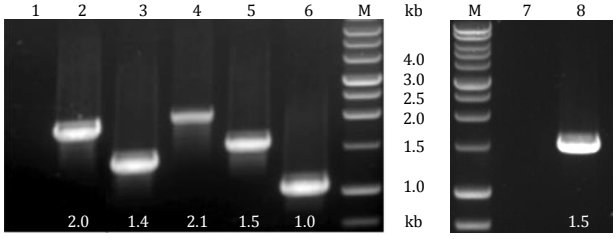


Figure 3.1. PCR amplification of glycosidase-encoding genes from *G. thermoglucosidasius* TM242 genomic DNA. Amplicons are shown on a 0.8% agarose gel corresponding to: (2) β-N-acetylhexosaminidase, (3) β-glucosidase, (4) β-xylosidase-1, (5) β-xylosidase-2, (6) Predicted glycoside hydrolase, (8) α-arabinofuranosidase, and (M) marker lanes containing a 1 kb ladder (Promega). Negative controls (no DNA) are shown in lanes 1 and 7. The expected sizes in kb for each of the PCR products are shown below each band.

3.3.3. Expression of TM242 glycosidases-encoding genes, and purification and characterisation of the recombinant enzymes

In order to evaluate the best culture conditions for *E. coli* harbouring pET-28a(+) with each TM242 glycosidase-encoding gene, a parallel analysis of protein expression and solubility was done by using two different rich media to support cell growth. Firstly, Terrific Broth (TB) was used, which provides higher density growth of *E. coli* cells and higher yield of plasmid DNA compared to LB broth, as it contains 20% more tryptone and 380% more yeast extract than LB. TB also contains phosphate buffer to maintain a physiological pH during exponential growth, and has an added 0.4% glycerol as an extra carbon source. This was compared with the auto-inducing medium Overnight Express™ Instant TB, which has been shown to produce cell mass and yields of target protein typically several-fold higher than obtained by conventional protocols using induction with IPTG (Grabski *et al.*, 2005; Studier, 2005). Successful expression of each TM242 glycosidase in soluble form was achieved with both culture media; the results obtained are individually presented below for each recombinant protein.

3.3.3.1. β-Xylosidase-1 (GH52)

The Protein Data Bank (www.pdb.org) is the single world-wide repository of structural data of biological macromolecules (Berman, 2000) with a comprehensive search tool that allows any protein to be compared with the structures available in the database. A PDB search with the amino acid sequence of the TM242 β-xylosidase-1 revealed a crystal structure of an A-type ATPase catalytic subunit A from *Pyrococcus horikoshii* OT3 (PDB entry: 1VDZ) as the most similar sequence in the database (Alignment: E-value: 7.53156; Score: 30.0314 bits (66); Identities: 23%; Positives: 44%; Gaps: 18%). This result was very interesting as it implies that, at the time of this search, there was no crystal structure available for this GH52 β–xylosidase, and even more promising, there were no crystal structures for any member of that family. As further structural studies might be useful for a better understanding of this enzyme, the recombinant TM242 β-xylosidase-1 was purified and fully characterised with the aim of obtaining the first GH52 crystal structure.

3.3.3.1.1. Protein expression and purification

The TM242 β-xylosidase-1 recombinant protein was correctly expressed in a soluble form using both media; however, the highest protein expression was achieved with TB medium and induction with 0.3 mM IPTG at OD_{600} =1 for 5 h. The cells were

then harvested, resuspended and lysed. The cell extract was purified by a two-step FPLC procedure to ensure the high purity as this protein was needed for structural studies (shown in Chapter 4). The results from the optimised first step of purification are shown in Figure 3.2.

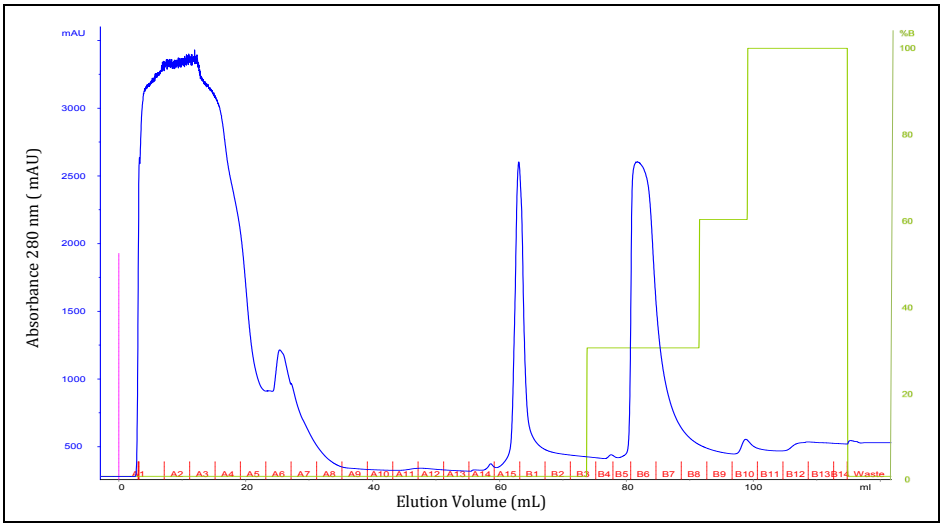


Figure 3.2. IMAC purification of β-xylosidase-1 expressed in *E. coli* BL21(DE3). Chromatography of a cell extract was carried out on a HisTrap HP 5mL column. The blue line corresponds to the A_{280} readings and the green line to the imidazole gradient. The fraction numbers are shown in red and the vertical pink dashed line corresponds to the sample injection. The column was washed with 10 CV of loading buffer (50 mM Tris–HCl pH 7.4, 500 mM NaCl), followed by a 4 CV wash with the same buffer containing 60 mM imidazole. The recombinant protein was finally eluted with 30% elution buffer (50 mM Tris–HCl pH 8.0, 300 mM NaCl, 500 mM imidazole) and the column was finally washed with a gradient of 60% and 100% of the same buffer. Fractions containing β-xylosidase activity were pooled and concentrated prior to gel filtration.

The fractions obtained from this first step purification were assayed with 10 mM of the artificial substrate *p*NP-β-D-xylopyranoside. Fractions B6-B10 presented activity indicating the elution of the recombinant β-xylosidase enzyme at a concentration of 200 mM imidazole. The fractions were visualised on a SDS-PAGE gel shown in Figure 3.3.

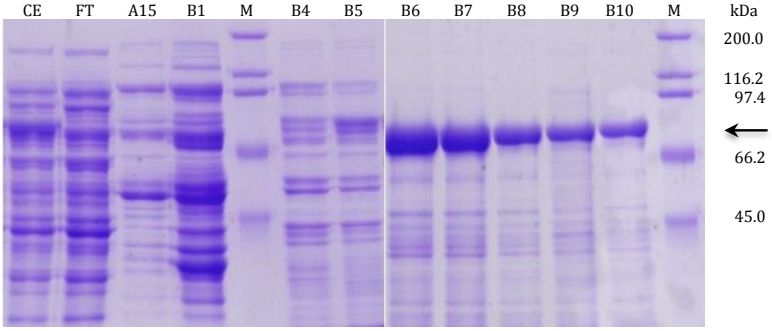


Figure 3.3. SDS-PAGE gel of the β-xylosidase-1 IMAC purification. Lanes are labelled by the fraction number shown in Figure 3.2. The loaded sample is shown in the lane labelled CE (cell extract) and the flow through (fractions A1-A14) is shown in lane labelled FT. The standard protein marker is shown in the lanes labelled M, and β-xylosidase-1 is highlighted with an arrow (predicted M_r = 82.1 kDa) in lanes B6-10.

As can be seen from the SDS-PAGE gel, even though the majority of the native *E. coli* proteins were not retained by the HisTrap column, some of them co-eluted with the β-xylosidase-1. For this reason a second step of purification was required to obtain high purity protein for the crystallisation trials. The fractions showing β-xylosidase activity (B6-B10) were pooled, concentrated and loaded onto a Superdex 200 10/300 GL size-exclusion column (GE Healthcare). The results from the optimised second-step purification are shown in Figure 3.4.

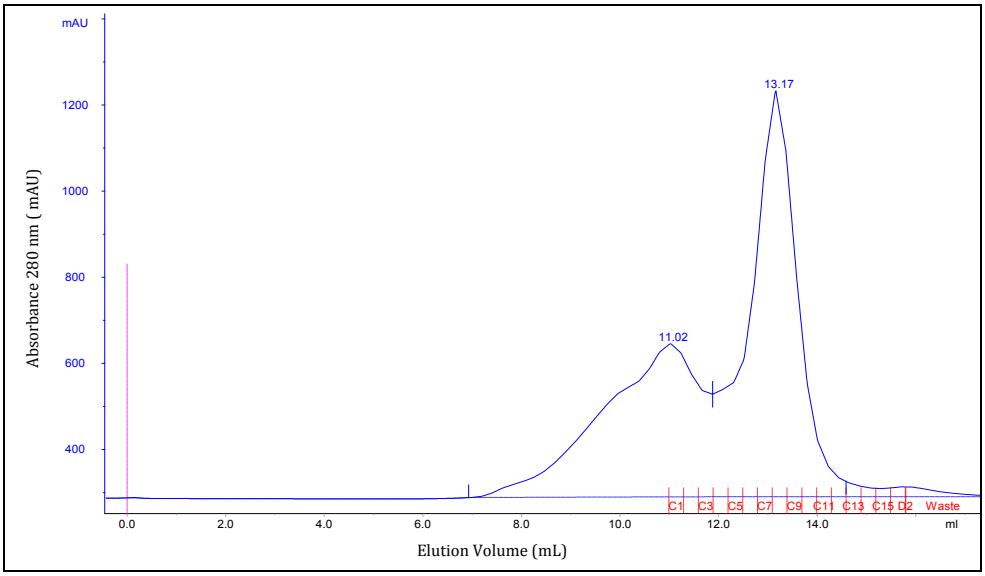


Figure 3.4. SEC purification of TM242 β-xylosidase-1. Chromatography of the concentrated enzyme (after a first step of IMAC purification) was carried out on a Superdex 200 10/300 GL size-exclusion column. The blue line corresponds to the A_{280} readings and the number above each peak corresponds to the elution volume integrated by UNICORN 5.0 software. The fractions numbers are shown in red and the vertical pink dashed line corresponds to the sample injection. The column was washed with 50 mM Tris-HCl buffer, pH 8.0, containing 150 mM NaCl.

The chromatograph shows two distinctive peaks that were assayed with 10 mM of *p*NP-β-D-xylopyranoside. No activity was found for fractions C1-C4 or C15-onwards, whereas fraction C5-C12 presented activity, indicating the elution of the recombinant β-xylosidase enzyme. Fractions containing active enzyme were visualised on an SDS-PAGE gel as shown in Figure 3.5.

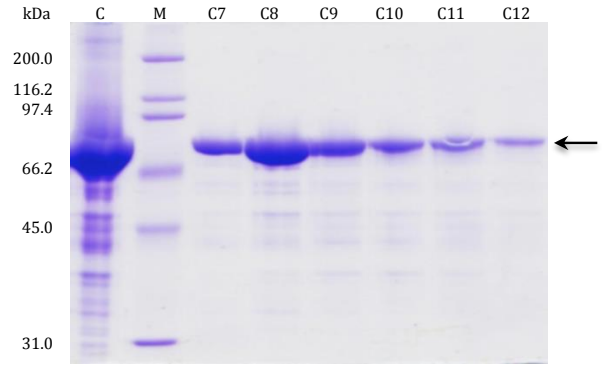


Figure 3.5. SDS-PAGE gel of the β-xylosidase-1 purified by SEC. Lanes are labelled by the fraction number shown in Figure 3.4. The load sample is shown in lane C (concentrated pooled fractions from IMAC). The standard protein marker (Biorad) is shown in lane M and β-xylosidase-1 is highlighted with an arrow (predicted M_r = 82.1 kDa).

Thus this second step of purification, was successful, in that a high degree of purification (~95%) has been achieved for the recombinant TM242 β-xylosidase-1. The protein sample was then suitable for crystallography, in which case it was concentrated to up to 20 mg/mL. Several other attempts of further purification after nickel affinity chromatography were made, including heat treatment (10-15 min at 60°C) and ion-exchange chromatography (both anion and cation using the GE Healthcare columns HiTrapQ at pH 8 and HiTrap SP at pH 6, respectively) based on the predicted pI of 5.62 obtained with ProtParam tool (Gasteiger *et al.*, 2003). However, size exclusion chromatography was the superior step of the β-xylosidase-1 purification process and a summary table of the two- step purification process is shown below.

Table 3.3. Summary of the β-xylosidase-1 two step purification process.

Enzyme sample	Total Protein* (mg)	Total Activity° (U)	Specific Activity (U mg^{-1})	Recovery (%)^
Cell extract	944	182	0.2	100
IMAC	37	145	3.9	80
SEC	24	96	4.0	40

*Protein concentration was determined by Bradford Assay. °Activity assays were performed at 65°C, pH 7 using 10 mM *p*NP-β-D-xylopyranoside as substrate. ^Only selected fractions were transferred from the activity peak after each chromatographic step, thus reducing the recorded yield. One unit (U) of enzyme activity is 1 µmol of *p*-nitrophenol per minute under assay conditions.

From a plot of the relative mobility versus the logarithm of the relative molecular mass of the standard proteins from the broad range markers (Biorad), the size of the β-xylosidase-1 was determined to be 81.8 kDa, which agrees with the expected size of 82.1 kDa calculated by the ProtParam tool (Gasteiger *et al.*, 2003). Size-exclusion chromatography with standard proteins of known molecular masses (Sigma) indicates that the β-xylosidase-1 elution volume corresponds to a molecular mass of 168 KDa, indicating that the oligomeric state of the enzyme is a dimer.

3.3.3.1.2. Protein characterisation

β-xylosidase-1 activity was routinely assayed with its artificial substrate *p*NP-β-D-xylopyranoside and it was characterised in terms of enzyme stability, thermoactivity, thermostability, kinetic parameters and substrate specificity. The enzyme stability after a 1 in 20 dilution in McIlvaine's buffer, pH 7.0, was evaluated and it was found that two weeks post dilution, stored at 4°C, the enzyme remained more than 75% active. The temperature optimum for this enzyme was determined to be 65°C under the specific conditions assayed (Figure 3.6), at which temperature the half-life ($T_{0.5}$) was 50 min as measured by its time-dependent irreversible thermal inactivation shown in Figure 3.7.

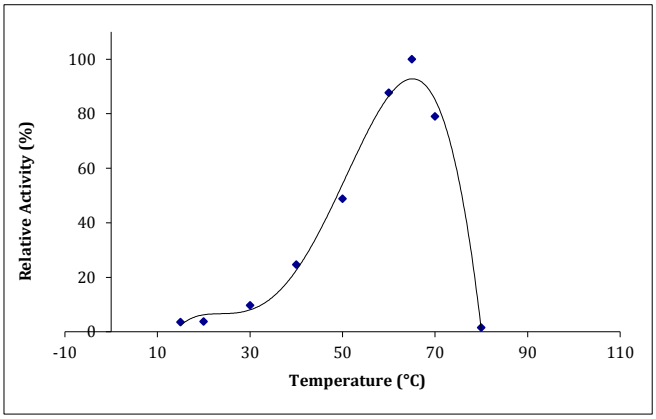


Figure 3.6. β-xylosidase-1 activity against temperature. The activity was expressed as percentage of the activity at 65°C.

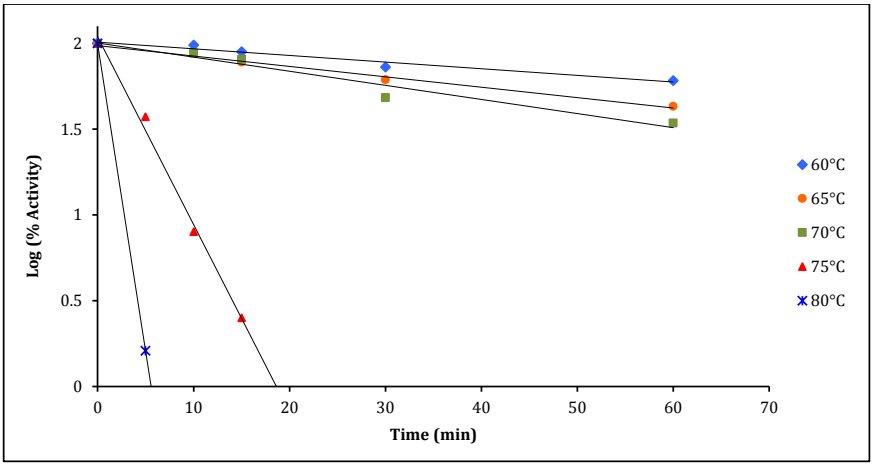


Figure 3.7. Thermal inactivation profile of β-xylosidase-1. Samples were heated at various temperatures for increasing lengths of time and then cooled prior to assaying the residual activity under optimal conditions. The enzyme activities of the heated samples are related to that of the non-heated enzyme.

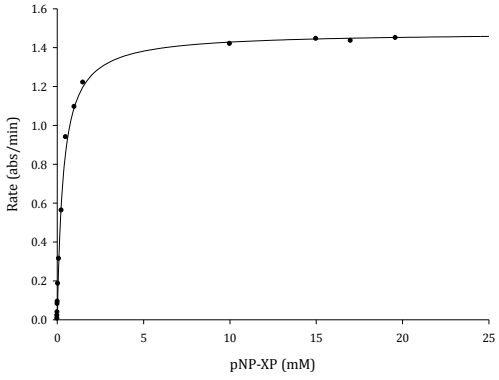
These thermostability assays showed the recombinant β-xylosidase-1 enzyme to be relatively stable between temperatures of 60-70°C over a 60-minute period, as might be expected for an enzyme from *G. thermoglucosidasius* growing optimally at 60°C. These results are also comparable with the ones from other family GH52 β-xylosidases from closely related microorganisms: *Geobacillus pallidus* (Quintero *et al.*, 2007), *Geobacillus stearothermophilus* 21 (Nanmori *et al.*, 1990) and T-6 (Bravman *et al.*, 2003), *Geobacillus thermodenitrificans* TSAA1 (Anand *et al.*, 2013) and *Paenibacillus sp.* DG-22 (Lee *et al.*, 2007).

The kinetic parameters were determined using the Enzyme Kinetics module in SigmaPlot 12 software based on the Wilkinson method which, together with the method described by Eisenthal & Cornish-Bowden (1974) (the direct linear plot), has been described to be the best to fit the Michaelis-Menten equation (Atkins & Nimmo, 1975). Data are also displayed as a Hanes-Woolf plot ([S]/v *versus* [S]), linearity confirming that the enzyme obeyed Michaelis-Menten kinetics. In addition to the specific substrate *p*NP-β-D-xylopyranoside (*p*NP-XP), the enzyme activity and its kinetic parameters were also evaluated using other *p*-nitrophenol-linked glycosides as substrates (*p*NP-β-D-glucopyranoside (*p*NP-GP), *p*NP-α–L-arabinopyranoside (*p*NP-AP), *p*NP-α-L-arabinofuranoside (*p*NP-AF)) to determine the enzyme substrate specificity. The results are shown in Table 3.4 and Figure 3.8.

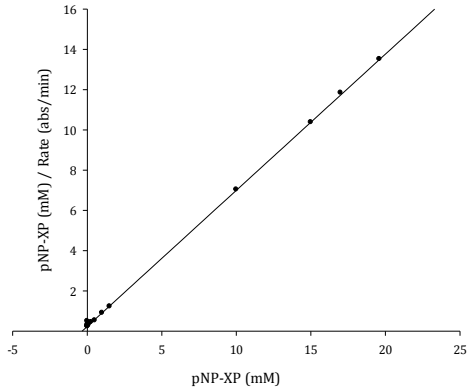
Table 3.4. Kinetic parameters of the recombinant β-xylosidase-1. Enzyme assays were carried out at 65°C. *V*max and K_M values were calculated using SigmaPlot 12 software. Standard Errors on all values are <10%.

Artificial substrates	**K_M (mM)**	***V*max (μmol *p*NP/min/mg)**	**k_{cat} (s^{-1})**	**k_{cat}/K_M ($mM^{-1}\,s^{-1}$)**
*p*NP- β-D-xylopyranoside	0.34	40.5	55.4	163
*p*NP-α-L-arabinofuranoside	0.8	0.3	0.4	0.5
*p*NP-α-L-arabinopyranoside	27.4	7.6	10.4	0.4
*p*NP- β-D-glucopyranoside	12.8	1.3	1.8	0.1

a)



b)



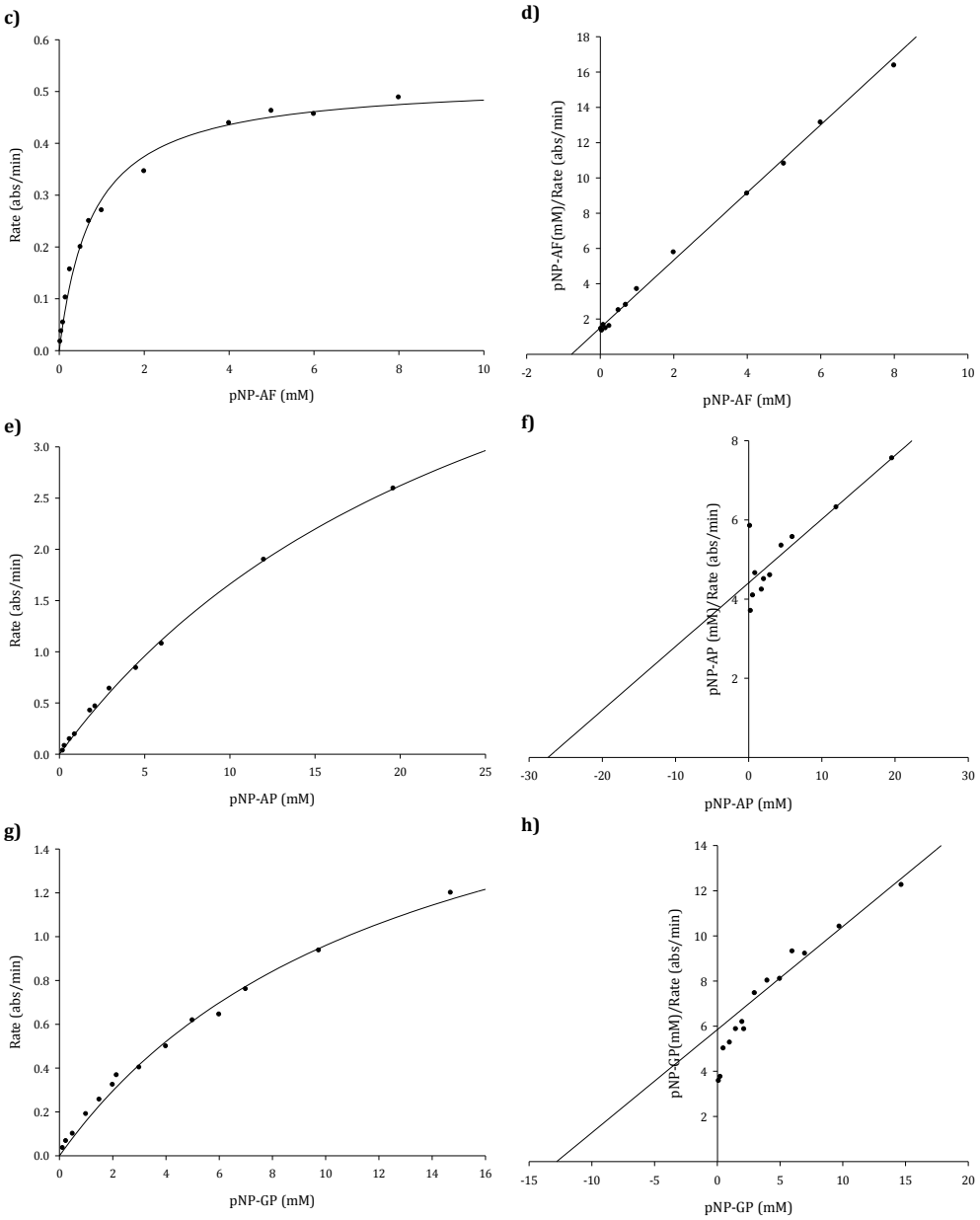


Figure 3.8. β-Xylosidase-1 kinetics. To determine the enzyme's substrate specificity, it was assayed with different concentrations of the substrates *p*NP-β-D-xylopyranoside (a,b), *p*NP-α-L-arabinofuranoside (c,d), *p*NP-α-L-arabinopyranoside (e,f), and *p*NP-β-D-glucopyranoside (g,h). The enzymatic reaction velocity is given in A_{410}/min. Left: Michaelis-Menten plot of the reaction velocity (v) as a function of the substrate concentration [S]. Right: Hanes-Woolf plot ([S]/v *versus* [S]).

Michaelis-Menten kinetics were observed with the substrates *p*NP-β-D-xylopyranoside and *p*NP-α-L-arabinofuranoside and it is clear that the enzyme has a strong preference for *p*NP-XP, although there was also significant activity with the *p*NP_α-L-arabinopyranoside derivative, albeit with a very high K_M. Furthermore, the data obtained are very similar to the kinetics of the GH52 β-xylosidase enzyme XynB2 from *G. stearothermophilus* T-6 (Bravman *et al.*, 2003) with which this enzyme shares

86% sequence identity, has a catalytic efficiency with *p*NP-β-D-xylopyranoside of 140 mM^{-1} s^{-1} and very similar K_M values for *p*NP-GP, *p*NP-AP and *p*NP-AF (12, 35, 0.71 mM respectively). The turnover number (k_{cat}), which represents the number of substrate molecules transformed to product (*p*NP) per enzyme molecule per second, was far superior for β-xylosidase-1 than XynB2 (at least 3-fold higher) with all the artificial substrates tested.

In order to assess if this enzyme was indeed a true β-xylosidase, the enzyme activity with its natural substrates, xylobiose and xylotriose, was investigated in different ways: Thin Layer Chromatography (TLC), inhibition assays and coupled enzyme assays. TLC was performed by the BSc Biochemistry student Victoria Cronin during her final year project to analyse qualitatively the hydrolysis of xylobiose catalysed by β-xylosidase-1. The results showed an almost immediate breakdown of xylobiose by the enzyme as xylose was detected after 30 s incubation with β-xylosidase-1 and xylobiose was not observed at any of the incubation points (Cronin, 2013).

Secondly, inhibition assays were carried out with *p*NP-XP in the presence of xylobiose (at different concentrations ranging from 0 to 3 mM). If xylobiose is a substrate of the enzyme, it will act as a competitive inhibitor with respective to the artificial substrate being used. The results shown in Figure 3.9 indicate that the natural substrate does indeed act as an inhibitor, indicating that it at least binds to the enzyme. However, unexpectedly, the inhibition does not appear to be strictly competitive.

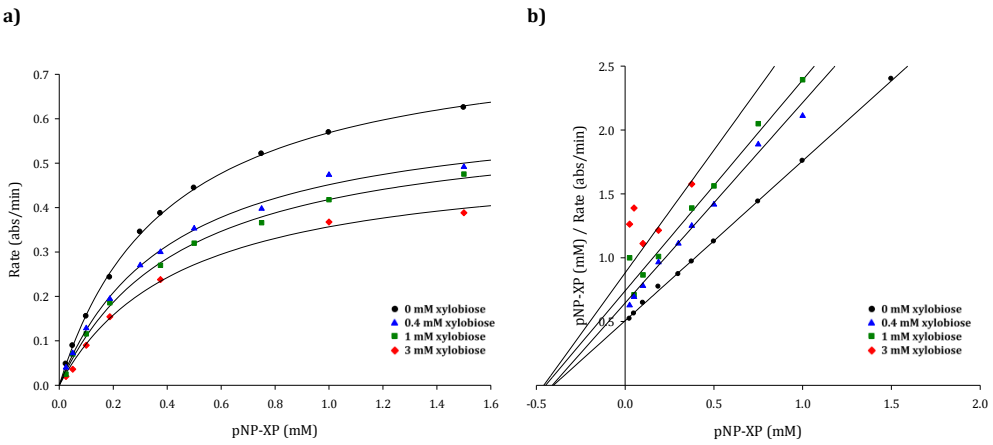


Figure 3.9. Effect of xylobiose on β-xylosidase activity. The enzyme was assayed with different concentrations of the substrate *p*NP-β-D-xylopyranoside at varying concentrations of xylobiose (0-3 mM). The enzymatic reaction velocity is given in A_{410}/min. (a) Michaelis-Menten plot of the reaction velocity (v) as a function of the substrate concentration [S]. (b) Hanes Woolf plot ([S]/v *versus* [S]), where parallel lines suggest competitive inhibition.

Treating the inhibition as competitive, the xylobiose inhibition constant, K_i, was determined to be 3.1 mM from the x-intercept of a secondary plot of K_M^{app} *versus* xylobiose concentration (Figure 3.10).

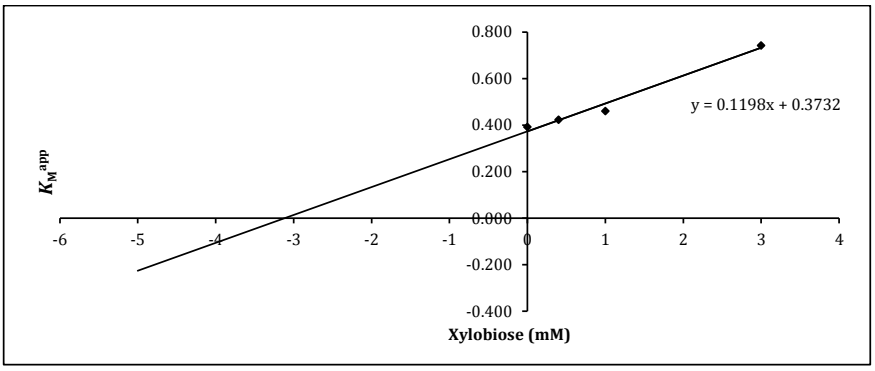


Figure 3.10. Secondary plot for competitive inhibition with xylobiose.

Finally, in a continuous spectrophotometric assay performed with the enzyme xylose dehydrogenase (β-XDH) (Megazyme), β-xylosidase-1 showed excellent activity with xylobiose and xylotriose (Table 3.5 and Figure 5.11) while no activity was found with xylan, thus confirming the enzyme's identity as a true β-xylosidase. On account of the mesophilic nature of the commercial β-XDH, the enzyme assays with the natural substrate could not be carried out at 65°C. However, from the data obtained at 45°C and 65°C with *p*NP-XP, the activity of the enzyme with xylobiose at 45°C (24.5 µmol *p*NP/min/mg) would equate to approximately 100 µmol *p*NP/min/mg at 65°C and a k_{cat} = 135 min^{-1}. Similar adjustments to the values with xylotriose give 68 µmol *p*NP/min/mg at 65°C and k_{cat} = 92 min^{-1}.

Table 3.5. Catalytic activity and kinetic parameters of the recombinant β-xylosidase-1. Enzyme assays were carried out at 45°C. *V*max and K_M values were calculated using SigmaPlot 12. Standard errors on all values are <10%. (-) indicates no activity could be detected.

Natural substrates	K_M (mM)	*V*max (µmol *p*NP/min/mg)	k_{cat} (s^{-1})	k_{cat}/K_M ($mM^{-1}\,s^{-1}$)
Xylobiose	2.4	24.5	33.5	13.9
Xylotriose	3.5	16.9	23.1	6.6
Xylan	-	-	-	-

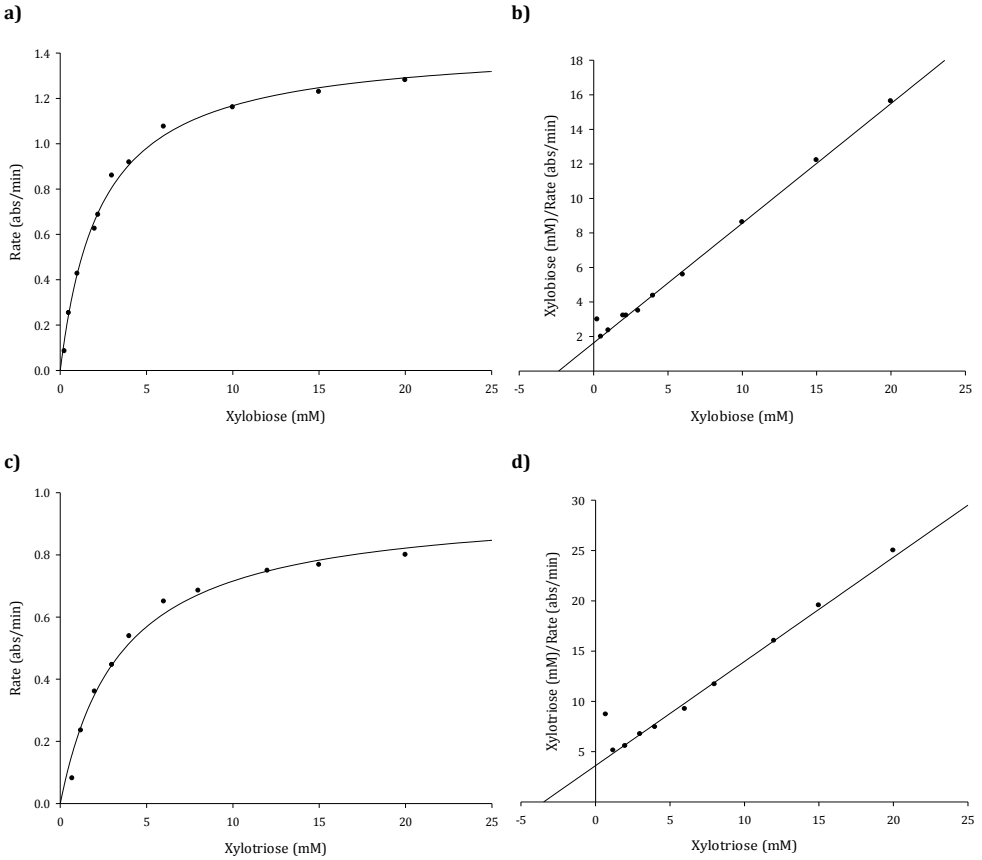


Figure 3.11. β-xylosidase-1 kinetics. To determine the enzyme's substrate specificity, it was assayed with different concentrations of its natural substrates, xylobiose (a,b) and xylotriose (c,d). The enzymatic reaction velocity is given in A_{340}/min. Left: Michaelis-Menten plot of the reaction velocity (v) as a function of the substrate concentration [S]. Right: Hanes-Woolf plot ([S]/v *versus* [S]).

In principle, the assays with the natural substrates could have been run at higher temperatures using a stopped-assay approach, incubating the β-xylosidase with xylobiose at 65°C, quenching the reaction at defined time-points and then measuring the xylose produced with the mesophilic coupling enzyme after each sample was cooled. However, continuous assays are easier and more accurate than discontinuous assays, a suitable stopping reagent or condition would have to be found to inactivate the xylosidase, and then samples would have to be treated again to provide conditions permitting the coupling enzyme to work. All this would have introduced potential errors in the assay. Furthermore, the K_i determined from the inhibition assays performed previously match the K_M determined for xylobiose using the coupled assay at 45°C, supporting the evidence that the kinetics reported at 45°C do accurately reflect the normal behaviour of the enzyme with the natural substrate, albeit at the expected reduced rate due to the lower temperature.

The above enzyme characterisation, together with the crystal structure presented in Chapter 4, are part of an original paper that has been published in *Acta Crystallographica Section D* (Appendix III).

3.3.3.2. β-Xylosidase-2 (GH43)

The TM242 β-xylosidase-2 recombinant protein was expressed in a soluble form using both media; however, the highest protein expression was achieved with the Overnight Express™ medium. The cells were then harvested, resuspended and lysed, and the enzyme was purified from the soluble cell extract by small-scale IMAC purification. An SDS-PAGE gel with each fraction is presented in Figure 3.12. From comparison with the standard proteins from the broad range marker (Biorad), the size of the β-xylosidase-2 was determined to be 65.2 kDa, which agrees with the expected size of 64.5 kDa calculated by the ProtParam tool (Gasteiger *et al.*, 2003).

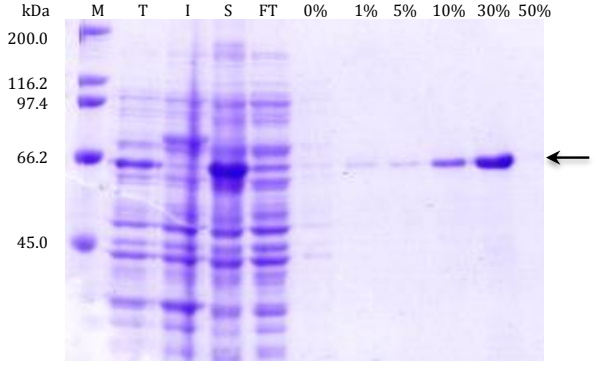


Figure 3.12. SDS-PAGE gel of the β-xylosidase-2 purified by IMAC. Lane T corresponds to the total cell extract, I is the insoluble fraction of the cell extract, S is the soluble fraction of the cell extract, and FT is the flow through after loading the soluble fraction onto the IMAC column. The following lanes are labelled by the percentage of His-elute buffer used to elute proteins from the column. The standard protein markers are shown in lane M and β-xylosidase-2 is highlighted with an arrow (predicted M_r = 64.5 kDa).

β-xylosidase-2 activity was routinely assayed with its articial substrate, *p*NP-β-D-xylopyranoside, and it was characterised in terms of enzyme stability, thermoactivity, thermostability, kinetics parameters and substrate specificity. The enzyme stability after a 1 in 20 dilution in McIlvaine's buffer, pH 7.0 was evaluated and it was found that, after one week post dilution stored at 4°C, the enzyme remained more than 85% active. The temperature optimum for the purified enzyme was determined to be 65°C under the specific conditions assayed (Figure 3.13), at which temperature the half-life ($T_{0.5}$) was 70 min as measured by its time-dependent irreversible thermal inactivation shown in Figure 3.14. These thermostability assays

showed the recombinant β-xylosidase-2 enzyme to be relatively stable between temperatures of 60-70°C over a 60-minute period.

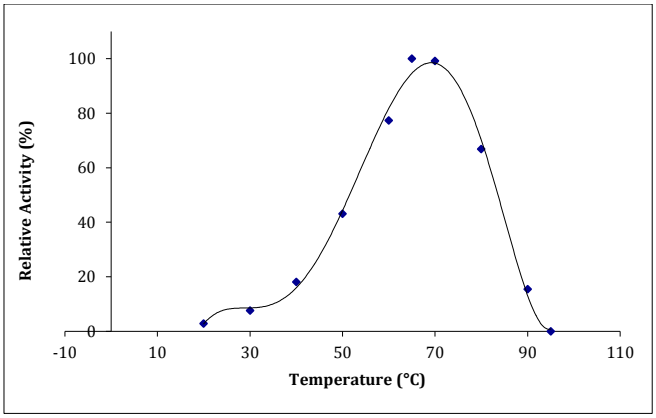


Figure 3.13. Temperature dependence of β-xylosidase-2 activity. The activity was assayed at temperatures between 20 and 95^{0}C and is expressed as a percentage of the activity at 65°C.

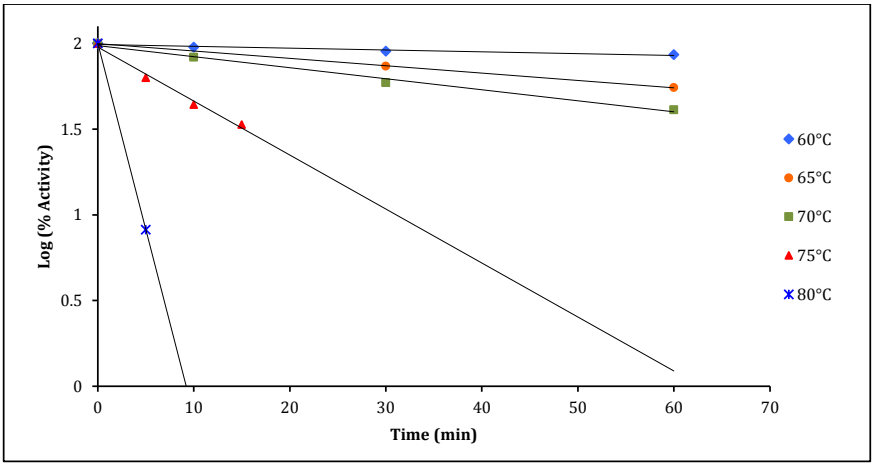


Figure 3.14. Thermal inactivation profile of β-xylosidase-2. Samples were heated at various temperatures for increasing lengths of time and then cooled prior to assaying the residual activity under optimal conditions. The enzyme activities of the heated samples are related to that of the non-heated enzyme.

The kinetic parameters were determined with SigmaPlot 12 for the specific substrate *p*NP-β-D-xylopyranoside as well as *p*NP-α–L-arabinofuranoside (experimental data obtained by Victoria Cronin, BSc Biochemistry project student) and the results are shown in Table 3.7 and Figure 3.15. Very low activity was found with the substrates *p*NP-β-D-arabinopyranoside and *p*NP-β-D-glucopyranoside at the concentrations tested and, due to their limited solubility, the kinetics parameters for these substrates could not be determined; however, it can be suggested that if the enzyme does hydrolyse them, it will do so with a very high K_M.

Table 3.6. Catalytic activity and kinetic parameters of the recombinant β-xylosidase-2. Enzyme assays were carried out at 65^{0}C. V_{max} and K_M values were calculated using SigmaPlot 12.

Artificial substrates	K_M (mM)	*V*max (µmol *p*NP/min/mg)	k_{cat} (s^{-1})	k_{cat}/K_M ($mM^{-1}\,s^{-1}$)
*p*NP- β-D-xylopyranoside	0.13 (± 0.02)	1.13 (± 0.007)	1.22 (± 0.007)	9.35 (± 1.44)
*p*NP-α-L-arabinofuranoside	1.15 (± 0.24)	0.15 (± 0.01)	0.16 (± 0.001)	0.14 (± 0.03))

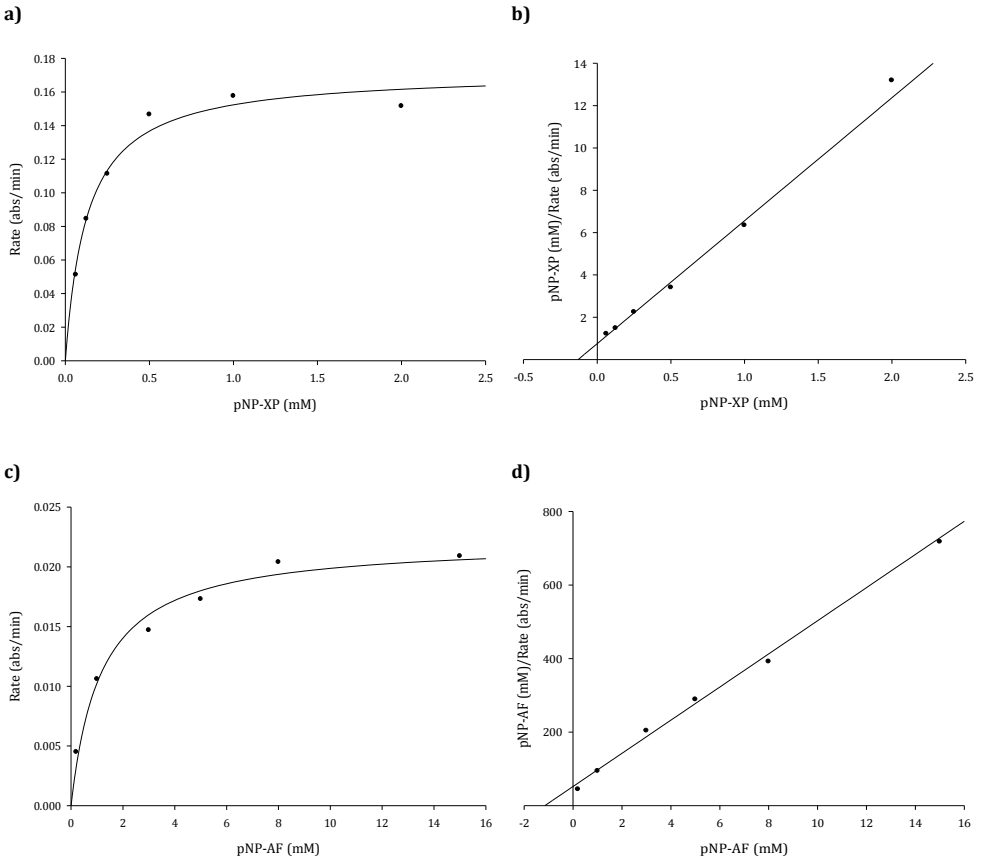


Figure 3.15. β-xylosidase-2 kinetics. To determine the enzyme's substrate specificity, it was assayed with different concentrations of the substrates *p*NP-β-D-xylopyranoside (a,b) and *p*NP-α-L-arabinofuranoside (c,d). The enzymatic reaction velocity is given in A_{410}/min. Left: Michaelis Menten plot of the reaction velocity (v) as a function of the substrate concentration [S]. Right: Hanes-Woolf plot ([S]/v *versus* [S]).

Michaelis-Menten kinetics were observed for both substrates tested, with the higher activity against *p*NP-XP as expected (7.5- fold that of activity with *p*NP-AF). β-xylosidase-2 seems to catalyse the hydrolysis of the artificial substrates with much lower catalytic efficiency in comparison with β-xylosidase-1 (17-fold lower with *p*NP-XP and 3.6 fold lower with *p*NP-AF), even though several enzymes from GH family 43 are known to display bifunctionality (Jordan & Li, 2007; Yang *et al.*, 2014), and in the ERGO database β-xylosidase-2 was annotated as a β-xylosidase/

α-L-arabinofuranosidase. However, β-xylosidase-1 appears to be superior not only in the hydrolysis of *p*NP-XP but also of *p*NP-AF with k_{cat} values 45 and 2.5 times higher than β-xylosidase-2, respectively.

A PDB search with the amino acid sequence of the TM242 β-xylosidase-2 revealed the GH43 β-xylosidase, XynB3, from *G. stearothermophilus* T-6 (PDB entry: 2EXH; structure obtained at 1.88 Å resolution) to have the most similar sequence in the database (Alignment: E-value: 0.0; Score: 964.526 bits (2492); Identities: 85%; Positives: 91%; Gaps: 0%). The XynB3 crystal structure, as described by Brüx *et al.* (2006), is a tetramer made up of two tight dimers (Figure 3.16). Each of these dimers consists of two monomers that are aligned antiparallel to each other and each monomeric subunit is organised in two domains, linked with an 11 residue linker: an N-terminal five-bladed β–propeller catalytic domain (residues 1–318), which is common to all GH43 members, and a β-sandwich C-terminal domain (residues 330 to 535) consisting of two sheets of 5 antiparallel β-strands (Brüx *et al.*, 2006).

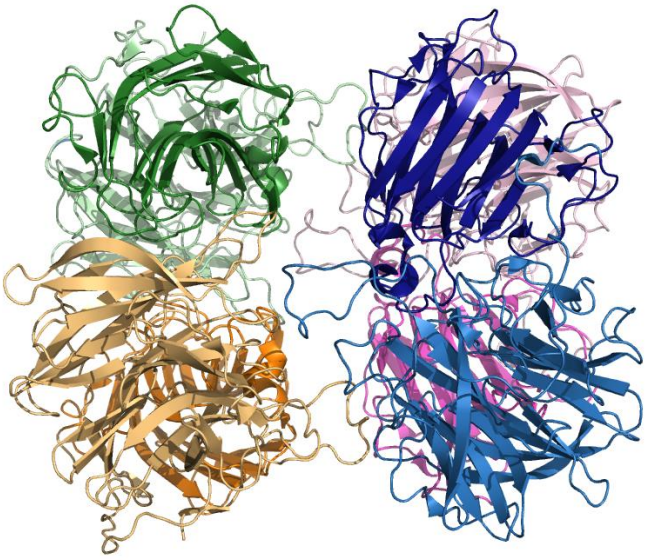


Figure 3.16. Cartoon representation of GH43 β-xylosidase (XynB3) from *Geobacillus stearothermophilus* (PDB entry: 2EXH). The protein structure is a tetramer made up of two dimers of two monomers aligned antiparallel to each other. Each monomeric subunit consist of an N terminal β-propeller and a C- terminal β-sandwich (Brüx *et al.*, 2006). Each monomer is coloured differently, with the N-terminal domains shown in lighter colours and C-terminal domains shown in darker colours.

Although a CDD search using the β-sandwich domain of XynB3 did not retrieve any sequences of carbohydrate binding modules, the authors suggested it is possible that this C-terminal domain formerly functioned as a CBM but lost this function during

evolution (Brüx *et al.*, 2006). In XynB3 this domain is involved in important inter-monomer interactions that lead to the observed oligomeric arrangement of the protein, and a small part of it contributes to the overall configuration of the substrate binding (Brüx *et al.*, 2006).

The catalytic active site of XynB3 is located at one end of the funnel-shaped cavity of the β-propeller domain that is formed as a result of the β-sheet packing (Figure 3.17). Like other GH43 members, three carboxylic residues are critical for the catalytic activity of XynB3: Asp15 (the general base), Glu187 (the general acid) and Asp128, while the amino acids Ser30, Trp74, Ala75, His249, Arg288, Thr207 and Phe506 are the ones involved in the interaction with the substrate (Brüx *et al.*, 2006). The active site possesses a pocket topology that is mainly constructed from β-propeller domain residues; the catalytic domain residues form a cleft that is blocked on one side by a loop originating from the β-sandwich domain (residues 494–507). This loop closes the cleft on one side restricting the possible length of the incoming substrate (*i.e.* the number of xylose units that can enter the active site), resulting in an exo-enzyme that releases only single xylose units (Brüx *et al.*, 2006).

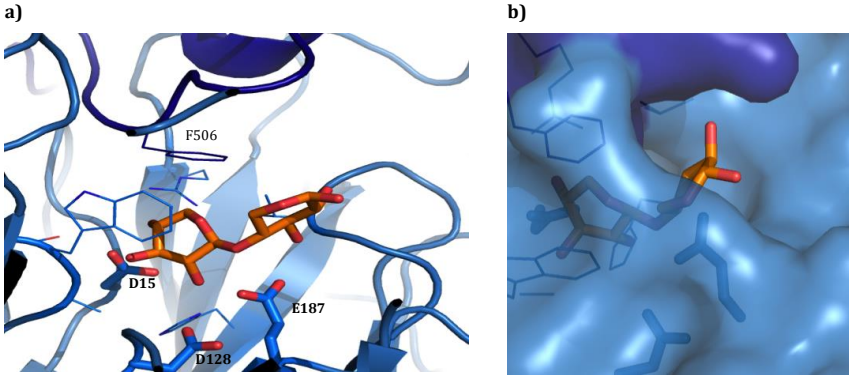


Figure 3.17. XynB3 active-site cleft active-site cleft with the xylobiose substrate shown in stick form (PDB entry: 2EXJ). (a) Residues forming the cleft are shown in stick form and the carboxylic catalytic residues D15, E187 and D128 (G128 in the crystal structure) are shown in thicker lines. F506 is the only residue from the β-sandwich domain, which is is shown in a darker colour. (b) A surface representation with 20% transparency.

A sequence alignment of both β-xylosidases is shown in Figure 3.18, and it can be seen that all the residues described to be crucial for XynB3 enzyme activity are fully conserved in the TM242 β-xylosidase-2, indicating that the crystal structure deposited in the PDB provides enough comparable information for structurally-based engineering of this enzyme. The potential for rational design to engineer the specificity of a GH43 enzyme has been shown by McKee *et al.* (2012), where the alteration of the active site

of the α-L-arabinofuranosidase (HiAXHd3) from *Humicola insolens* by the mutation of Tyr166 to Ala166 (residue not conserved in GH43 β-xylosidases), generated an enzyme that displays xylanase activity while retaining its α-arabinofuranosidase activity. The authors suggest that the five-bladed β-propeller fold displayed by GH43 enzymes provides a structural platform that can be harnessed to bind a range of different sugars and to catalyse the hydrolysis of glycosidic bonds through distinct modes of action in order to generate multifunctional enzymes (McKee *et al.*, 2012).

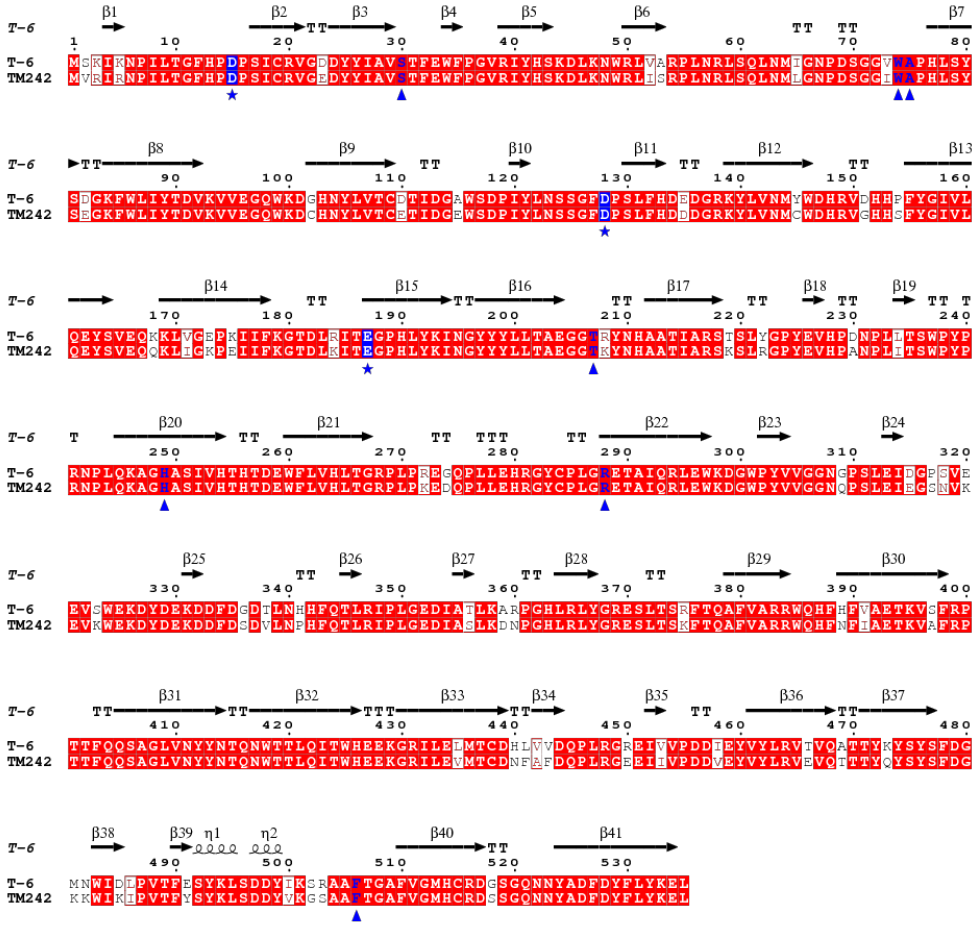


Figure 3.18. Structure-based sequence alignment of GH43 β-xylosidase from *G. stearothermophilus* T-6 and β-xylosidase-2 from *G. thermoglucosidasius* TM242, produced using Clustal Omega and displayed using ESPript 3.06 (Gouet *et al.*, 1999). Identical residues are shown in white on a red background, while conservative amino acid changes are shown in red. Amino acids involved in the interaction with substrate (Brüx *et al.*, 2006) are highlighted in blue (the catalytic residues are indicated by blue stars, and the others by blue triangles beneath the alignment). Secondary structure of *G. stearothermophilus* T-6 β-xylosidases, XynB3, is shown above the alignment with β-strands as arrows and helices as coils.

3.3.3.3. α-Arabinofuranosidase (GH51)

The TM242 α-arabinofuranosidase recombinant protein was expressed in a soluble form using both media; however, the highest protein expression was achieved with TB medium and induction with 0.5 mM IPTG at OD_{600} =1 for 3.5 h. The cells were then harvested, resuspended and lysed, and the enzyme was purified from the soluble cell extract by small-scale IMAC purification. An SDS-PAGE gel with each fraction is presented in Figure 3.19. From comparison with the standard proteins from the broad range marker (Biorad), the size of the α-arabinofuranosidase was determined to be 62.4 kDa, which agrees with the expected size of 60.2 kDa calculated by the ProtParam tool (Gasteiger *et al.*, 2003).

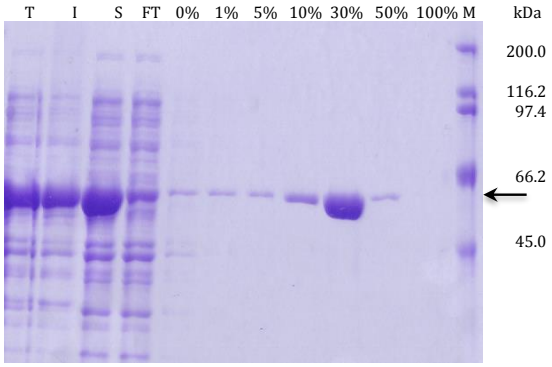


Figure 3.19. SDS-PAGE gel of the α-arabinofuranosidase IMAC purification. Lane T corresponds to the total cell extract, I is the insoluble fraction of the cell extract, S is the soluble fraction of the cell extract, and FT is the flow through after loading the soluble fraction onto the IMAC column. The following lanes are labelled by the percentage of His-elute buffer used to elute proteins from the column. The standard protein markers are shown in lane M and α-arabinofuranosidase is highlighted with an arrow (predicted Mr = 60.2 kDa).

α-Arabinofuranosidase activity was routinely assayed with its articial substrate *p*NP-β-D-arabinofuranoside and it was characterised in terms of enzyme stability, thermoactivity, thermostability, kinetic parameters and substrate specificity by the BSc Biochemistry student, Emma Goffin, during her final-year project. The enzyme stability after a 1 in 20 dilution in McIlvaine's buffer, pH 7.0, was evaluated and it was found that two weeks post dilution, stored at 4°C, the enzyme remained more than 95% active. Furthermore, some of the α-arabinofuranosidase was found in the insoluble fraction (see Figure 3.19), but this could be resolubilised with 8M urea after which the enzyme was assayed found to be catalytically active (Goffin, 2014).

The temperature optimum for the purified enzyme (from the soluble cell extract) was determined to be 80°C under the specific conditions assayed (Figure 3.20) and, as measured by its time-dependent irreversible thermal inactivation (Figure 3.21), the

enzyme had a half-life of approximately 70 min at 75°C, while incubating at 80°C near abolished enzyme activity within less than 10 minutes. These thermostability assays showed the recombinant α-arabinofuranosidase enzyme to be relatively stable between temperatures of 60-75°C over a 60-minute period, which is consistent with other GH51 α-arabinofuranosidases from the closely related microorganisms *Geobacillus caldoxylolyticus* TK4 (Canakci *et al.*, 2007), *Geobacillus stearothermophilus* T-6 (Gilead & Shoham, 1995), and *Thermobacillus xylanilyticus* (Debeche *et al.*, 2007).

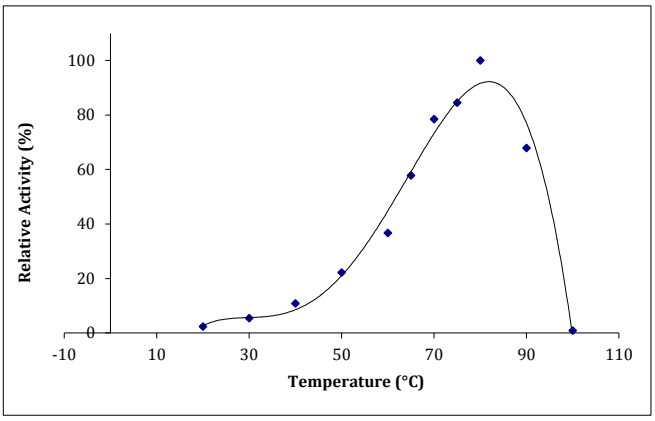


Figure 3.20. Temperature dependence of α-arabinofuranosidase activity. The activity was assayed at temperatures between 20 and 95^{0}C and is expressed as a percentage of the activity at 80°C.

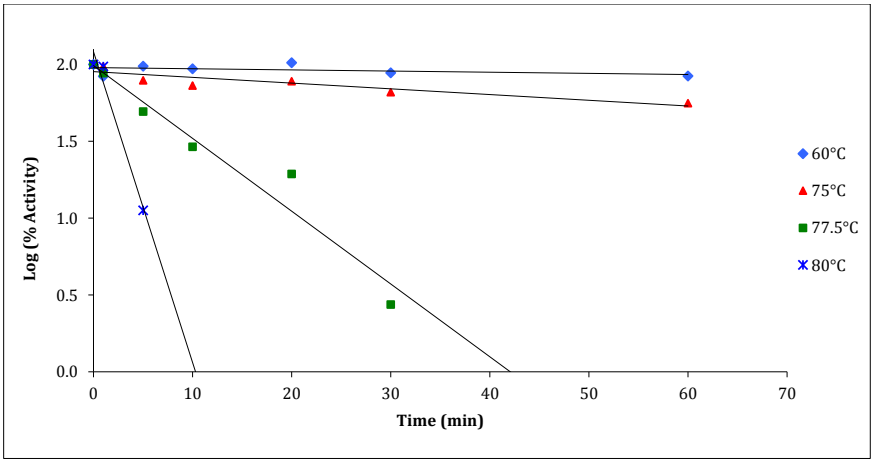


Figure 3.21. Thermal inactivation profile of α-arabinofuranosidase. Samples were heated at various temperatures for increasing lengths of time and then cooled prior to assaying the residual activity under optimal conditions. The enzyme activities of the heated samples are related to that of the non-heated enzyme.

The kinetic parameters were determined with SigmaPlot 12 for the specific substrate *p*NP-α-L-arabinofuranoside as well as *p*NP-β-D-xylopyranoside and *p*NP-β-D-glucopyranoside (experimental data obtained by Emma Goffin, BSc Biochemistry project student) and the results are shown in Figure 3.22 and Table 3.8.

Very low activity was found with the substrate *p*NP-α-L-arabinopyranoside at the concentrations tested and, due to its limited solubility, the kinetics parameters for this substrate could not be determined but it can be suggested that if the enzyme can hydrolyse this substrate, it will do so with a very high K_M.

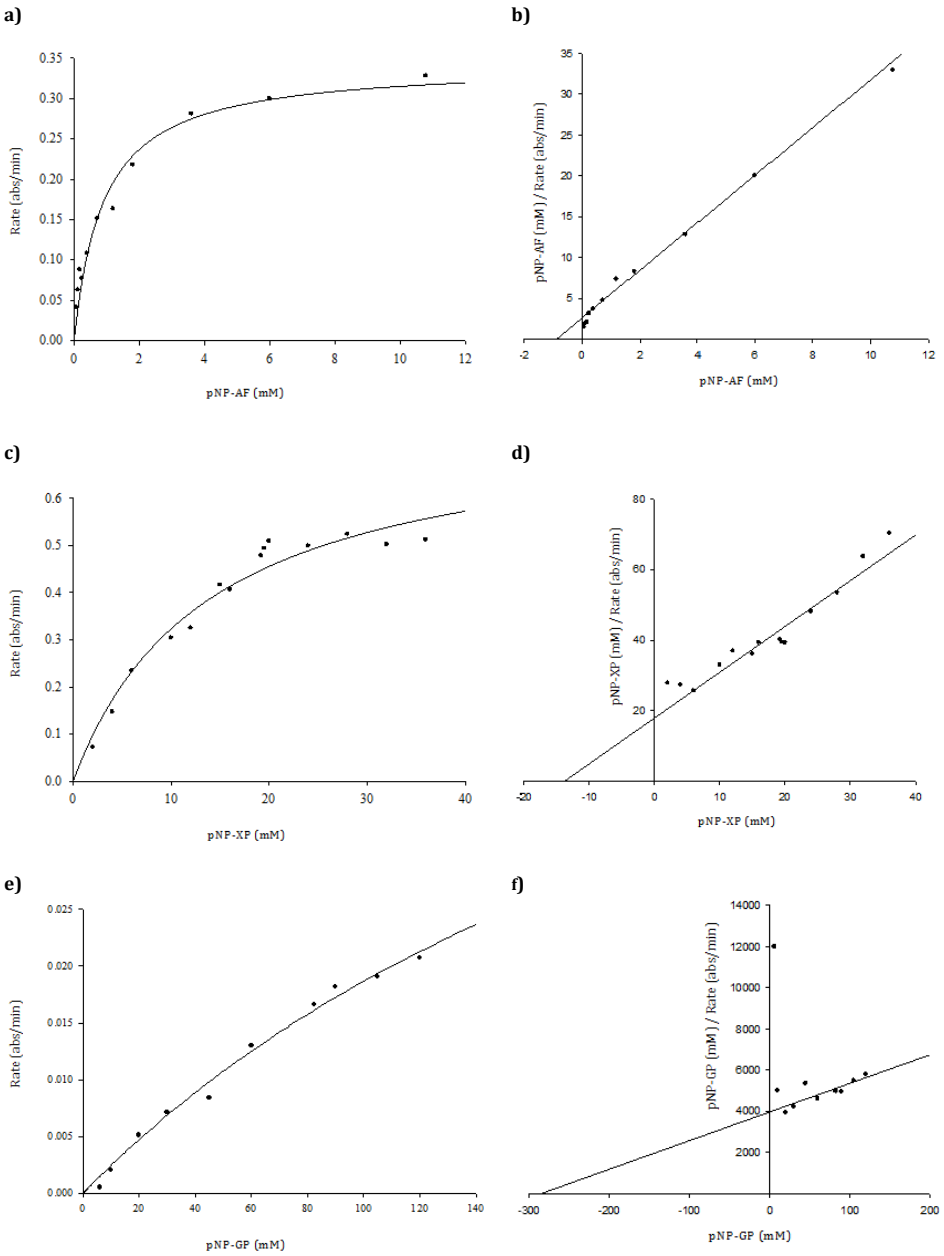


Figure 3.22. α-Arabinofuranosidase kinetics. To determine the enzyme's substrate specificity, it was assayed with different concentrations of the substrates *p*NP-α-L-arabinofuranoside (a,b), *p*NP-β-D-xylopyranoside (c,d) and *p*NP-β-D-glucopyranoside (e,f). The enzymatic reaction velocity is given in A_{410}/min. Left: Michaelis-Menten plot of the reaction velocity (v) as a function of the substrate concentration [S]. Right: Hanes-Woolf plot ([S]/v *versus* [S]).

Table 3.7. Kinetic parameters of the recombinant α-arabinofuranosidase. Enzyme assays were carried out at 60^{0}C. V_{max} and K_M values were calculated using SigmaPlot 12.

Artificial substrates	K_M (mM)	Vmax (μmol *p*NP/min/mg)	k_{cat} (s^{-1})	k_{cat}/K_M (mM^{-1} s^{-1})
*p*NP-α-L-arabinofuranoside	0.9 (± 0.1)	237.0 (± 11.6)	237 (± 11.6)	263 (± 32)
*p*NP- β-D-xylopyranoside	13.8 (± 2.6)	2.7 (± 0.2)	2.7 (± 0.2)	0.2 (± 0.55)
*p*NP- β-D-glucopyranoside	285.2 (± 96.9)	0.2 (±0.1)	0.2 (±0.1)	-

Michaelis-Menten kinetics were observed with all the three substrates tested and it is clear that the enzyme has a strong preference for *p*NP-AF, although there was also activity with *p*NP-XP and *p*NP-GP, albeit with higher K_M values (15- fold and 317- fold, respectively). Furthermore, the data obtained are comparable to the kinetics of the GH51 α-arabinofuranoside enzyme AbfA from *G. stearothermophilus* T-6 (Shallom *et al.*, 2002) with which shares 90% sequence identity. AbfA has a similar K_M (0.65 mM) to the TM242 enzyme, but a k_{cat} almost three times lower (87 s^{-1}); thus the catalytic efficiency with *p*NP-AP is 2-fold lower (130 mM^{-1} s^{-1}).

The enzyme has been described as a true α–L-arabinofuranosidase due its ability to hydrolyse the natural substrate arabinan (Goffin, 2014); further characterisation of this enzyme's catalytic activity with other arabinose-containing polymers such as sugar beet pulp is currently under investigation by Dr Charlotte Bennett (University of Bath) using High Performance Liquid Chromatography (HPLC) to measure the release of arabinose and other sugars by the action of the TM242 α-arabinofuranosidase.

A PDB search with the amino acid sequence of the TM242 α-arabinofuranosidase revealed the GH51 α-arabinofuranosidase, AbfA, from *G. stearothermophilus* T-6 (PDB entry: 1PZ2; structure obtained at 2 Å resolution) to have the most similar sequence in the database (Alignment: E-value: 0.0; Score: 969.533 bits (205); Identities: 90%; Positives: 97%; Gaps: 0%). The AbfA crystal structure, as described by Hövel *et al.* (2003), is a homo-hexamer that can be considered as a trimer of dimers (Figure 3.23); each dimer consists of two monomers that are aligned antiparallel to each other and each monomeric subunit is organised in two domains: a catalytic N-terminal $(\beta/\alpha)_8$ barrel (TIM barrel) domain (residues 20- 383), which is common to all GH51 members, and a C-terminal 12-stranded β-sandwich domain (residues 6-9 and 384-501) with a jelly-roll topology that has eight β-strands arranged in two closely packed sheets that are stabilised by hydrophobic interactions (Hövel *et al.*, 2003b).

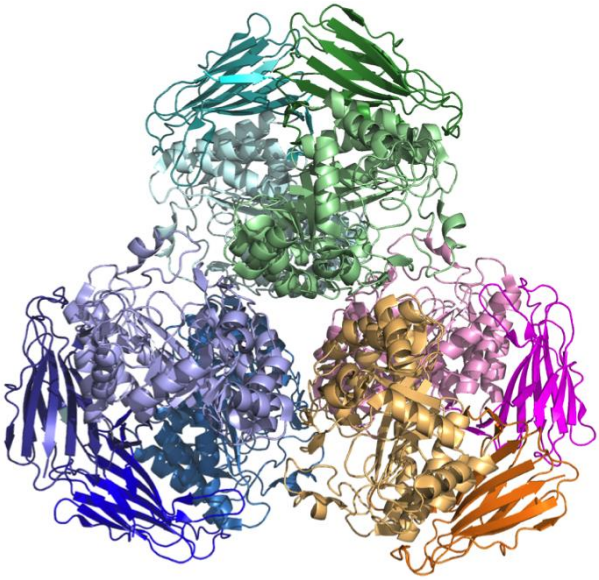


Figure 3.23. Cartoon representation of GH51 α-arabinofuranosidase (AbfA) from *Geobacillus stearothermophilus* (PDB entry: 1PZ2). The protein structure is a hexamer made up of a trimer of dimers of two monomers aligned antiparallel to each other. Each monomeric subunit consist of an N-terminal $(\beta/\alpha)_8$ barrel and a C-terminal β-sandwich (jelly roll) (Hövel *et al.*, 2003b). Each monomer is coloured differently, with the N-terminal domains shown in lighter colours and C-terminal domains shown in darker colours.

Although a CDD search of the β-sandwich domain of AbfA did not retrieve any sequences of carbohydrate binding modules, this domain shows structural similarity to cellulose binding domains and domain C of several α-amylases; the authors suggest it is possible that this C-terminal domain formerly functioned as a CBM but lost this function during evolution, and proposed that the jelly-roll domain might play a role stabilizing the TIM-barrel structure (Hövel *et al.*, 2003b).

The catalytic active site of AbfA is located at the C-terminal ends of β-strands 4 and 7 in the $(\beta/\alpha)_8$-barrel fold (Figure 3.24). Two carboxylic amino acids were identified as the catalytic residues: Glu175 (the acid/base) and Glu294 (the nucleophile) (Shallom *et al.*, 2002; Hövel *et al.*, 2003), while the amino acids Glu29, Arg69, Asn74, Asn174, His244, Tyr246 and Gln351 are responsible for substrate-binding interactions (Hövel *et al.*, 2003b).

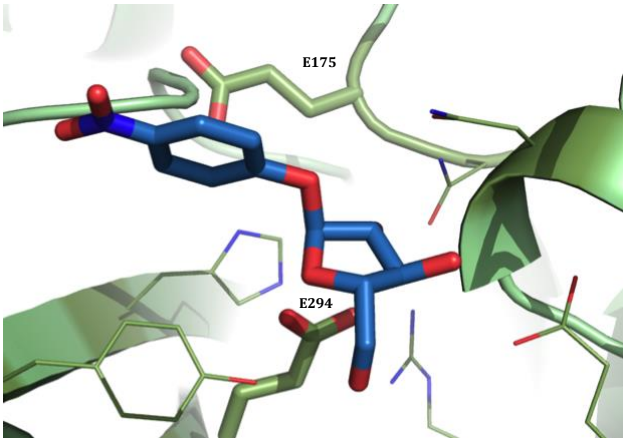


Figure 3.24. AbfA active-site cleft in complex with xylobiose (PDB entry: 1QW9). The substrate (*p*NP-AF) is shown in thick bonds with orange carbons. The residues forming the cleft are shown in stick form and the carboxylic catalytic residues E175 (A175 in the crystal structure) and E294 are shown in thicker lines.

The crystal structure reveals that the enzyme's active site binds arabinofuranosyl sugars tightly with relatively low distortion, and it has been shown that, due to D-xylopyranose and L-arabinofuranose sharing spatial similarity, the AbfA active site can also accommodate xylopyranosidic substrates but without the necessary distortion required for the efficient catalysis of six-membered rings (Wagschal *et al.*, 2009). This offers an explanation for the activity but lower specificity of GH51 α-arabinofuranosidases towards the xylopyranosidic substrates (Hövel *et al.*, 2003b). This is consistent with the kinetics data shown in Table 3.6, where the catalytic efficiency of the TM242 α-arabinofuranosidase against *p*NP-XP was 1315 times lower than against *p*NP-AF.

A sequence alignment of both α-arabinofuranosidases is shown in Figure 3.25, and it can be seen that all the residues described to be crucial for AbfA enzyme activity are fully conserved in the TM242 α-arabinofuranosidase, indicating that the crystal structure deposited in the PDB provides enough comparable information for structurally-based engineering of this enzyme.

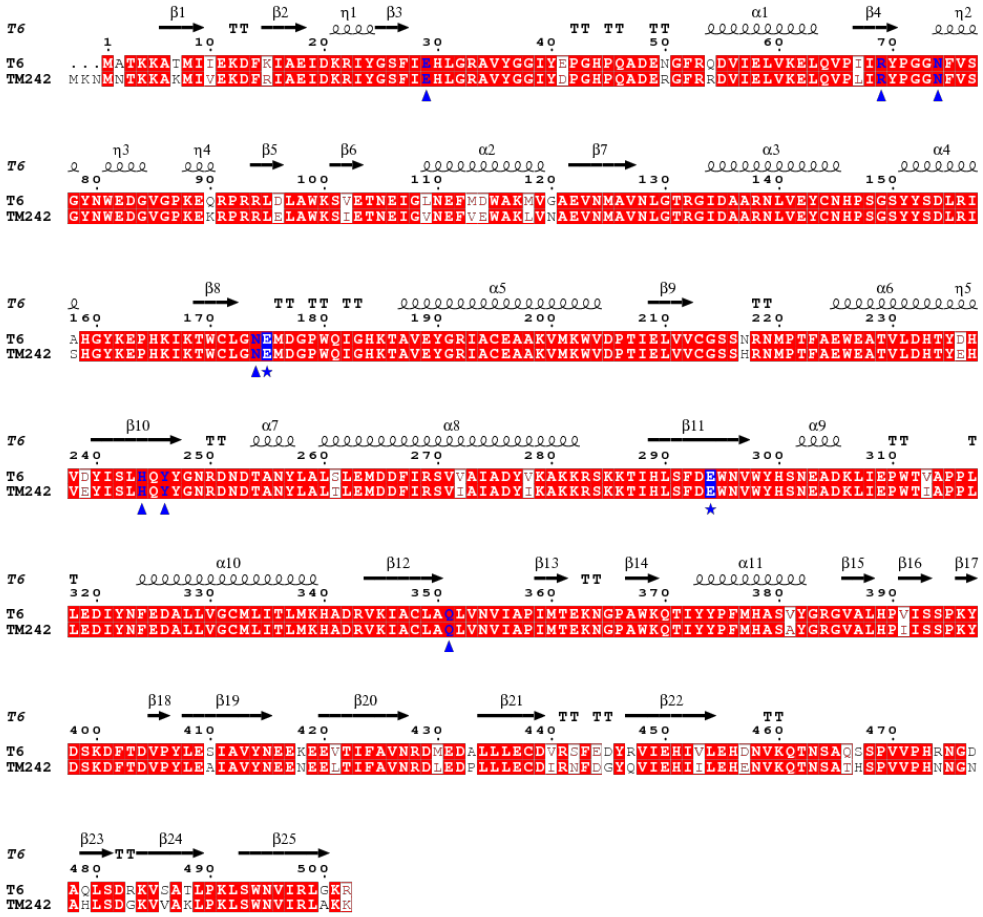


Figure 3.25. Structure-based sequence alignment of GH51 α-arabinofuranosidases from *G. stearothermophilus* T-6 and *G. thermoglucosidasius* TM242, produced using Clustal Omega and displayed using ESPript 3.06 (Gouet *et al.*, 1999). Identical residues are shown in white on a red background, while conservative amino acid changes are shown in red. Amino acids involved in the interaction with substrate (Hövel *et al.*, 2003b) are highlighted in blue (the catalytic residues are indicated by blue stars, and the others by blue triangles beneath the alignment). Secondary structure of *G. stearothermophilus* T-6 α-L-arabinofuranosidase, AbfA, is shown above the alignment with β-strands as arrows and helices as coils.

3.3.3.4. β-N-Acetylhexosaminidase (GH3)

The TM242 β-N-acetylhexosaminidase recombinant protein was expressed in a soluble form using both media; however, the highest protein expression was achieved with Overnight Express™ medium. The cells were then harvested, resuspended and lysed, and the enzyme was purified from the soluble cell extract by small-scale IMAC purification. An SDS-PAGE gel with each fraction is presented in Figure 3.26. From comparison with the standard proteins from the broad range marker (Biorad), the size of the β-N-acetylhexosaminidase was determined to be 73.3 kDa, which agrees with the expected size of 75.6 kDa calculated by the ProtParam tool (Gasteiger *et al.*, 2003).

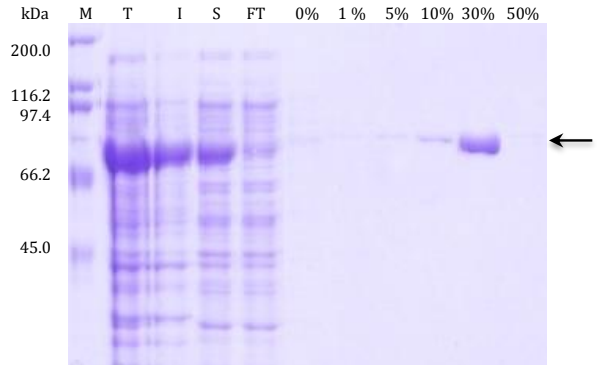


Figure 3.26. SDS-PAGE gel of the β-N-acetylhexosaminidase IMAC purification. Lane T corresponds to the total cell extract, I is the insoluble fraction of the cell extract, S is the soluble fraction of the cell extract, and FT is the flow through after loading the soluble fraction onto the IMAC column. The following lanes are labelled by the percentage of His-elute buffer used to elute proteins from the column. The standard protein markers are shown in lane M and β-N-acetylhexosaminidase is highlighted with an arrow (predicted Mr= 75.6 kDa).

β-N-Acetylhexosaminidase activity was routinely assayed with its articial substrate *p*NP-N-acetyl-β-D-glucosaminide and it was characterised in terms of enzyme stability, thermoactivity, thermostability, kinetics parameters and substrate specificity. The enzyme stability after a 1 in 20 dilution in McIlvaine's buffer, pH 7.0, was evaluated and it was found that two week's post dilution, stored at 4°C, the enzyme remained more than 90% active. The temperature optimum for the purified enzyme was determined to be 70°C under the specific conditions assayed (Figure 3.27), at which temperature the half-life ($T_{0.5}$) was 280 min as measured by its time-dependent irreversible thermal inactivation shown in Figure 3.28. These thermostability assays showed the recombinant β-N-acetylhexosaminidase enzyme to be relatively stable between temperatures of 60-80°C over a 60-minute period.

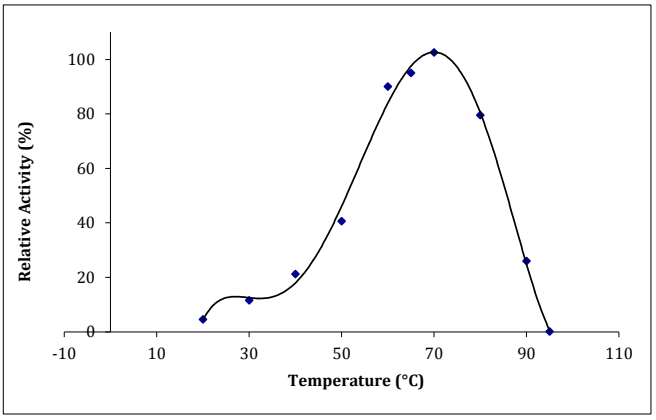


Figure 3.27. Temperature dependence of β-N-acetylhexosaminidase activity. The activity was assayed at temperatures between 20 and 95^{0}C and is expressed as a percentage of the activity at 70°C.

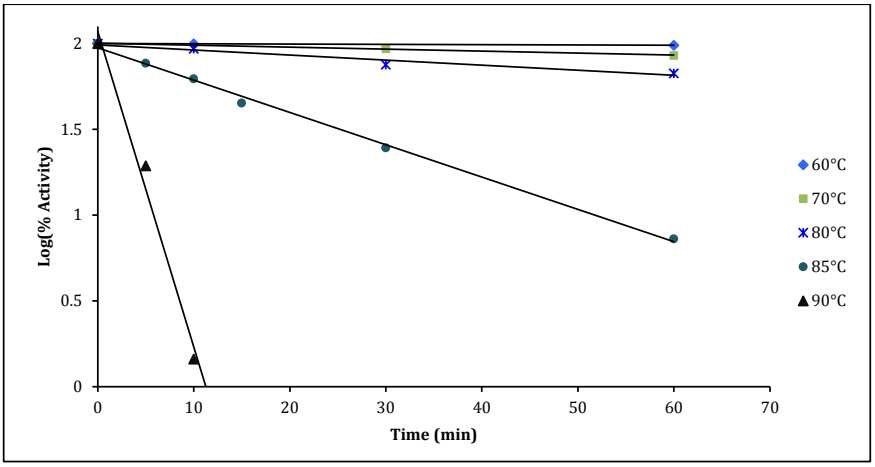


Figure 3.28. Thermal inactivation profile of β-N-acetylhexosaminidase. Samples were heated at various temperatures for increasing lengths of time and then cooled prior to assaying the residual activity under optimal conditions. The enzyme activities of the heated samples are related to that of the non-heated enzyme.

The kinetic parameters were determined with SigmaPlot 12 for the specific substrate *p*NP-N-acetyl-β-D-glucosaminide (*p*NP-NAG) and the results are shown in Figure 3.29 and Table 3.9. The enzyme displays Michaelis-Menten kinetics. No activity was found against *p*NP-GP, *p*NP-XP, *p*NP-AF and *p*NP-AP, thus confirming that the ERGO annotation of this enzyme is correct and effectively corresponds to a β-N-acetylhexosaminidase that is not associated with cellulose or hemicellulose hydrolysis.

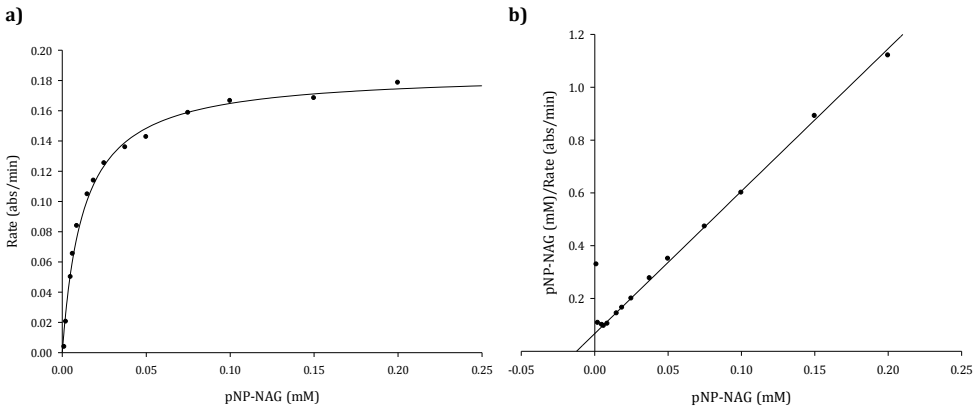


Figure 3.29. β-N-acetylhexosaminidase kinetics. The enzyme was assayed with *p*NP-N-acetyl-β-D-glucosaminide The enzymatic reaction velocity is given in A_{410}/min. Left: Michaelis-Menten plot of the reaction velocity (v) as a function of the substrate concentration [S]. Right: Hanes-Woolf plot ([S]/v *versus* [S]).

Table 3.8. Kinetic parameters of the recombinant β-N-acetylhexosaminidase. Enzyme assays were carried out at 70°C. V_{max} and K_M values were calculated using SigmaPlot 12.

Artificial substrates	K_M (μM)	*V*max (μmol *p*NP/min/mg)	k_{cat} (s^{-1})	k_{cat}/K_M ($mM^{-1}\ s^{-1}$)
*p*NP-N-acetyl-β-D-glucosaminide	12.4	8.1	10.2	823

A PDB search with the amino acid sequence of β-N-acetylhexosaminidase revealed the GH3 β-N-hexosaminidase, NagZ, from *Bacillus subtilis* (PDB entry: 3BMX; structure obtained at 1.4 Å resolution) to be the most similar sequence in the database (Alignment: E-value: 6.93815E-133; Score: 471.855 bits (1213); Identities: 45%; Positives: 62%; Gaps: 9%). The NagZ crystal structure described by Litzinger *et al.* (2010) is monomeric and consists of two separate domains: an N-terminal domain (residues 26–420) that adopts a $(\beta/\alpha)_8$-barrel fold and a C-terminal domain (residues 421–642) that displays an αβα-sandwich fold (Figure 3.31). The surface buried between the domains is indicative of an intimate association, although no part of the C-terminal domain comes into close contact with the substrate/inhibitor binding site on top of the $(\beta/\alpha)_8$-barrel domain.

Figure 3.30. Cartoon representation of GH3 β-N-hexosaminidase (NagZ) from *Bacillus subtilis* (PDB entry: 3BMX). The protein structure is a monomer consisting of an N terminal $(\beta/\alpha)_8$ barrel and a C-terminal αβα-sandwich (Litzinger *et al.*, 2010). The N-terminal domain is shown in lighter colour and the C-terminal a domain is shown in darker colour.

3.3.3.5. β-Glucosidase (GH1)

The TM242 β-glucosidase recombinant protein was expressed in a soluble form using both media; however, the highest protein expression was achieved with Overnight Express™ medium. The cells were then harvested, resuspended and lysed, and the enzyme was purified from the soluble cell extract by small-scale IMAC purification. An SDS-PAGE gel with each fraction is presented in Figure 3.32. From comparison with the standard proteins from the broad range marker (Biorad), the size of the β-glucosidase was determined to be 55.4 kDa, which agrees with the expected size of 57.3 kDa calculated by the ProtParam tool (Gasteiger *et al.*, 2003).

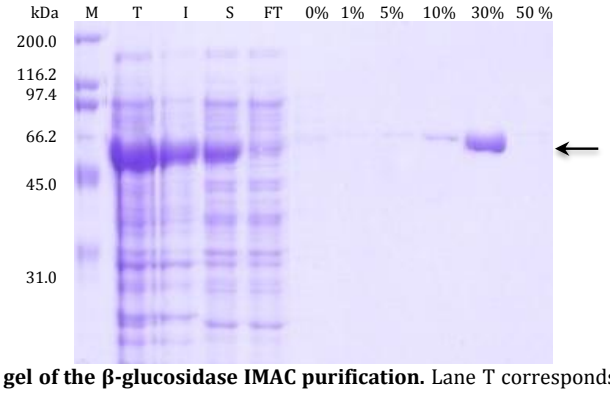


Figure 3.31. SDS-PAGE gel of the β-glucosidase IMAC purification. Lane T corresponds to the total cell extract, I is the insoluble fraction of the cell extract, S is the soluble fraction of the cell extract, and FT is the flow through after loading the soluble fraction onto the IMAC column. The following lanes are labelled by the percentage of His-elute buffer used to elute proteins from the column. The standard protein markers are shown in lane M and β-glucosidase is highlighted with an arrow (predicted Mr = 57.3 kDa).

The recombinant β-glucosidase was assayed with the artificial substrates *p*NP-GP, *p*NP-XP, *p*NP-AF, *p*NP-AP and *p*NP-NAG, but the enzyme was found to be inactive with all of them. It has been reported that a Tris molecule has been located in the active site of two different GH1 β-glucosidases crystal structures (TrBgl2 from *Trichoderma reesei* and NkBgl from the termite *Neotermes koshunensis*) interacting directly with the residues of the glycone-binding site and inhibiting the enzyme activities, and that a concentration of 2–10 mM Mn^{2+} significantly enhanced the hydrolase activity of the β-glucosidases by 150% to 230% (Jeng *et al.*, 2011). Therefore, the TM242 β-glucosidase was also purified with buffer containing HEPES instead of Tris and assayed in presence of different concentrations of the cation Mn^{2+}, but no improvement was observed and the recombinant protein remained inactive against the substrates tested.

A PDB search with the amino acid sequence of the TM242 β-glucosidase revealed a GH1 6-phospho-β-D-glucosidase, SPy1599, from *Streptococcus pyogenes* (PDB: 4B3K; structure obtained at 2.6 Å resolution) to be the most similar sequence in the database (Alignment: E-value: 2.31792E-129; Score: 459.914 bits (1182); Identities: 48%; Positives: 67%; Gaps: 2%). The SPy1599 crystal structure described by Stepper *et al.* (2013) is a classical $(\beta/\alpha)_8$ barrel fold (Figure 3.33).

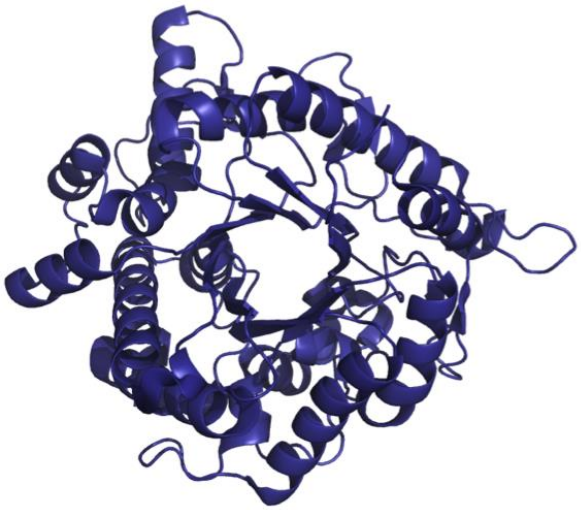


Figure 3.32. Cartoon representation of GH1 6-phospho-β-D-glucosidase (SPy1599) from *Streptococcus pyogenes* (PDB entry: 4B3K). The protein structure is monomeric with a $(\beta/\alpha)_8$ barrel fold (Stepper *et al.*, 2013).

Even though this enzyme was annotated in sequence depositions as a β–glucosidase, no such activity could be found by the authors; instead, three-dimensional structural overlaps with other enzymes of known function suggested that SPy1599 contains a phosphate-binding pocket in the active site and subsequent kinetic analysis showed that SPy1599 indeed possesses 6-phospho-β-glucosidase (EC 3.2.1.86) activity, confirming the misannotation (Stepper *et al.*, 2013).

Two carboxylic amino acids were identified as the SPy1599 catalytic residues: the Glu165 (the acid/base) located in the NEP motif at the end of β4, and Glu366 (the nucleophile) situated in the ENG motif at the end of β7, while the amino acids Arg74, His119, Asn164, Asn297, Tyr299 and Trp416 are the residues surrounding the active site (Stepper *et al.*, 2013). A sequence alignment of both proteins is shown in Figure 3.34, and it can be seen that, although their sequence identity is 48%, all the residues described to be crucial for SPy1599 enzyme activity are fully conserved in the TM242 β-glucosidase, suggesting its misannotation and potentially explaining the lack of

activity against the substrates tested. In order to confirm its identity as 6-phospho-β-glucosidase, enzyme assays with the specific *p*NP-β-D-glucopyranoside 6-phosphate are needed.

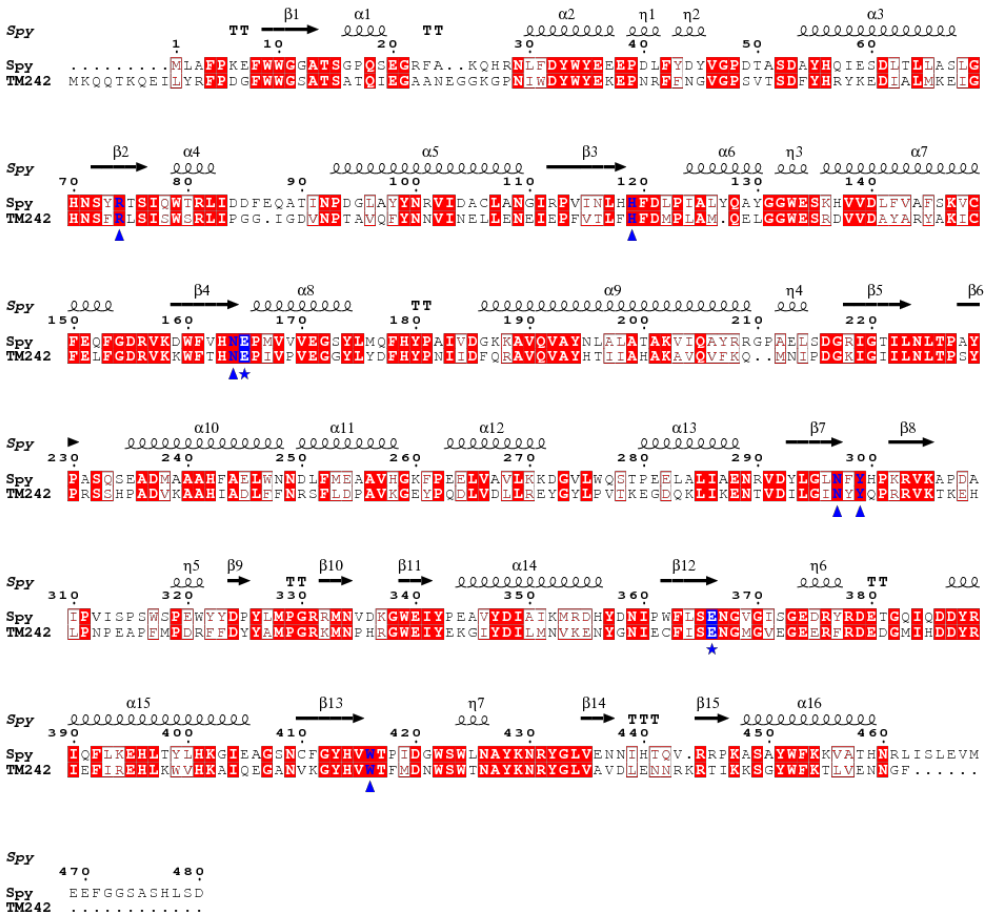


Figure 3.33. Structure-based sequence alignment of GH1 6-phospho-β-D-glucosidase from *Streptococcus pyogenes* and β-glucosidase from *G. thermoglucosidasius* TM242, produced using Clustal Omega and displayed using ESPript 3.06 (Gouet *et al.*, 1999). Identical residues are shown in white on a red background, while conservative amino acid changes are shown in red. Amino acids involved in the interaction with substrate (Stepper *et al.*, 2013) are highlighted in blue (the catalytic residues are indicated by blue stars, and the others by blue triangles beneath the alignment). Secondary structure of *Streptococcus pyogenes* 6-phospho-β-D-glucosidase, SPy1599, is shown above the alignment with β-strands as arrows and helices as coils.

3.3.3.6. Predicted glycoside hydrolase

The TM242 predicted glycoside hydrolase recombinant protein was expressed in a soluble form using both media; however, the highest protein expression was achieved with Overnight Express™ medium. The cells were then harvested, resuspended and lysed, and the enzyme was purified from the soluble cell extract by small-scale IMAC purification. An SDS-PAGE gel with each fraction is presented in Figure 3.35. From comparison with the standard proteins from the broad range marker (Biorad), the size of the predicted glycoside hydrolase was determined to be 42.2 kDa, which agrees with the expected size of 40.1 kDa calculated by the ProtParam tool (Gasteiger *et al.*, 2003).

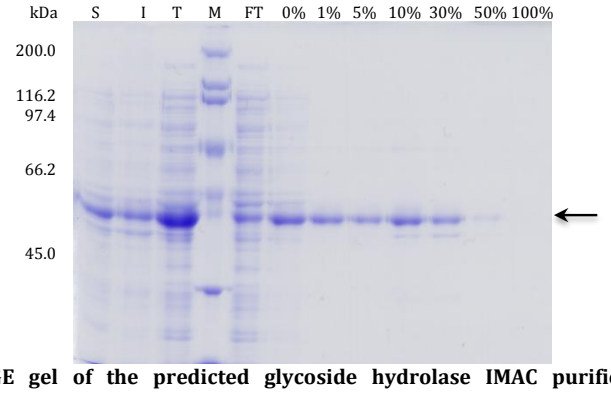


Figure 3.34. SDS-PAGE gel of the predicted glycoside hydrolase IMAC purification. Lane T corresponds to the total cell extract, I is the insoluble fraction of the cell extract, S is the soluble fraction of the cell extract, and FT is the flow through after loading the soluble fraction onto the IMAC column. The following lanes are labelled by the percentage of His-elute buffer used to elute proteins from the column. The standard protein markers are shown in lane M and the predicted glycoside hydrolase is highlighted with an arrow (predicted Mr = 40.1 kDa).

The recombinant predicted glycoside hydrolase was assayed with all the artificial substrates available (*p*NP-GP, *p*NP-XP, *p*NP-AF, *p*NP-AP and *p*NP-NAG), but the enzyme was found to be inactive with all of them.

A structural search of the PDB using the amino acid sequence of the TM242 predicted glycoside hydrolase revealed another predicted glycosidase (tm1225) from *Thermotoga maritima* msb8 (PDB entry: 1VKD; structure obtained at 2.1 Å resolution) to be the most similar sequence in the database (Alignment: E-value: 1.02217E-39; Score: 161.384 bits (407); Identities: 37%; Positives: 55%; Gaps: 4%). Even though this structure was deposited in the PDB on 2004, to date it has not been published. From the crystal structure deposited, it can be said that this protein is a homo-dimer with a five-bladed β-propeller fold (Figure 3.36), that is common to the GH families 32, 43, 62 and 68. All the other structures found to share sequence similarities to the

TM242 predicted glycoside hydrolase sequences are described as predicted, putative or hypothetical glycoside hydrolases. Therefore, further assays using other substrates are required in order to reveal the identity of this protein.

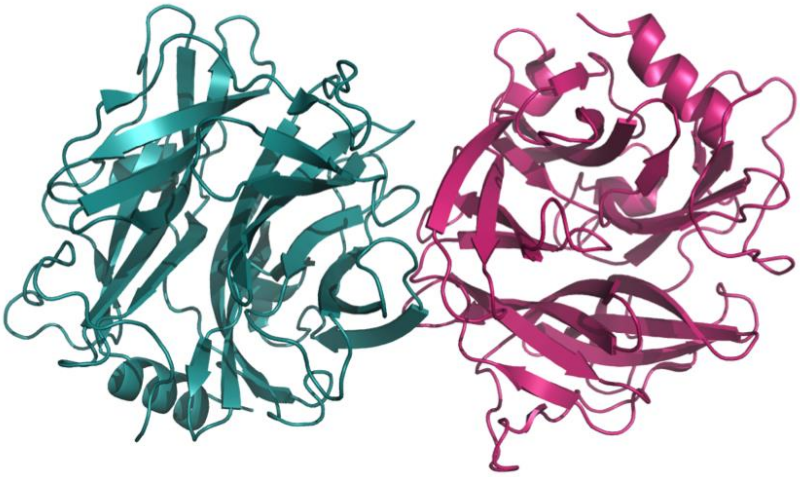


Figure 3.35. Cartoon representation of predicted glycosidase (tm1225) from *Thermotoga maritima* msb8 (PDB entry: 1VKD). The protein structure is a dimer with a five-bladed β–propeller fold. Each monomer is coloured differently.

3.4. CONCLUSIONS

Six different TM242 glycoside hydrolase gene sequences have been successfully cloned and expressed in *E. coli* BL21 using the pET-28a(+) vector, and the soluble recombinant proteins have been purified by nickel-affinity chromatography. In the case of β-xylosidase-1, this was a first step followed by size exclusion chromatography in order to prepare a protein sample suitable for crystallisation studies (described in Chapter 4).

Four of the recombinant proteins (β-xylosidase-1, β-xylosidase-2, α-arabinofuranosidase and β-N-acetylhexosaminidase) display enzymatic activity using *p*-nitrophenol-linked glycosides as substrates. Moreover, their substrate specificity, kinetic parameters, temperature optima and thermostabilities were determined. The enzymes had optimal temperatures expected for the thermophilic microorganism *G. thermoglucosidasius* TM242 and, even though expressed in a mesophilic host, they were relatively stable between temperatures of 60-70°C over a 60-minute period.

Further characterisation was carried out with β-xylosidase-1, assaying the enzyme with its natural substrates. In summary, the results obtained indicate that β-xylosidase-1 is able to hydrolyse short xylo-oligosaccharides and possesses far superior catalytic activity than β-xylosidase-2.

α-Arabinofuranosidase is very active and highly stable. However, β-N-acetylhexosaminidase is the most active and thermostable of the enzymes shown in this chapter, presenting the highest catalytic efficiency with its specific artificial substrate and enzymatic activity even after 60 minutes at 85°C.

The results obtained with the putative β-glucosidase suggest that the protein might be misannotated in ERGO, actually being a 6-phospho-β-D-glucosidase with no β-glucosidase activity. Finally the function of the TM242 predicted glycoside hydrolase remains unknown and the use of other substrates is required to reveal its true enzymic activity.

CHAPTER 4: Structural determination of β-xylosidase-1: the first crystal structure of a glycoside hydrolase from family GH52.

4.1. INTRODUCTION

As described in the previous chapter, the genes encoding six glycoside hydrolases from *Geobacillus thermoglucosidasius* TM242 have been cloned and expressed, and those recombinant enzymes shown to be active have been characterised. A search of the PDB database (Berman, 2000) using their amino acid sequences revealed that structures sharing 37- 85% sequence identity were available for five of these proteins, while the sequence of β-xylosidase-1 from glycoside hydrolase family 52 did not reveal any homologous structure. Moreover, an analysis of family GH52 in the CAZy database showed that no structure had been reported for any member of the family, indicating that structural studies on this enzyme and the information obtained from them could be of broad interest.

Protein structures are important to understand a protein's biological function and its physical and biochemical properties. In the case of enzymes, the structure unveils details about the position of the active site, substrate accessibility and potential inhibitors; in addition, the information about the amino acids involved in the catalysis and its interactions may enable a reaction mechanism to be proposed. Furthermore, structural information can also assist with the design and interpretation of experiments and it provides the basis for structurally-informed engineering of enzymes, which, as stated by Royer *et al.* (2010), reduces the time and effort that is required to improve the catalytic activity through directed evolution (using random mutagenesis) that often requires the screening of thousands of mutants.

Different methods can be used to determine the three-dimensional molecular structure of a protein, including X-ray crystallography, Nuclear Magnetic Resonance (NMR spectroscopy), cryo-electron microscopy and Small-Angle X-ray and Neutron Scattering (SAXS and SANS respectively). To date, most of the structures deposited in the PDB database have been determined using X-ray crystallography and, due to the size of the xylosidase enzyme, this was the method of choice during the pursuit of the β-xylosidase-1 structure.

4.1.1. X-ray crystallography

The fineness of detail that can be observed with any type of radiation is defined by its wavelength, the resolution limit being objects that are no closer than half a wavelength apart (Blow, 2002). For this reason, in order to image the structure of molecules in which bonded atoms are only about 1.5 Ångström (0.15 nm) apart, radiation with wavelength no longer than a couple of Ångström units (Å) is needed (Rhodes, 1993).

X-rays are electromagnetic waves that have wavelengths comparable to atomic dimensions that make them particularly useful for atomic structure determination. X-rays, with their very short wavelength, have very high energies and can be absorbed and scattered by matter. Individual molecules diffract X-rays very weakly and suffer rapid radiation damage, while crystals, as they contain many copies of a molecule, can amplify the signal and produce strong detectable X-ray beams (Rhodes, 1993).

Protein crystals constitute an ordered, three-dimensional periodic arrangement of protein molecules held together by non-covalent interactions (Drenth, 1994). As proteins are large and irregular, the molecular contacts are not uniform within the crystal and they contain a high solvent content, which makes them very fragile (Blow, 2002). Crystals of protein can grow by slow, controlled precipitation from aqueous solution under supersaturated conditions that do not denature the protein but, instead, favour crystal nucleation and growth. The exact crystallisation conditions (*e.g.* buffer pH, salts, precipitant, presence of solvents, additional ions or sugars, temperature, protein concentration and purity) cannot be predicted, as they vary depending on the properties of each protein (as does the time it will take the crystal to grow). However, it is believed that generally the higher the concentration and purity of a protein, the better to promote crystallisation (Drenth, 1994).

X-ray crystallography determines the three dimensional structure of a protein from the diffraction pattern produced by exposing a suitable protein crystal to an intense beam of X-rays. By analysing the diffraction pattern, and the spacing of the characteristic array of spots (reflections) obtained, it is possible to estimate how the molecules are arranged in the crystal (Goodsell, 2010). Each of these reflections produced by the beam of electromagnetic waves (X-rays) is characterised by its amplitude and phase, and can be described by a Fourier series called the structure factor equation:

$$\boldsymbol{F}_{hkl} = F_{hkl} e^{(i\alpha_{hkl})} = \sum_j f_j \, e^{[2\pi i(hx_j + ky_j + lz_j)]}$$

where $\boldsymbol{F}_{hkl}$ is the structure factor that describes the diffracted waves from scattering plane *hkl* and α_{hkl} is the phase of the diffracted beam; the summation is over all atomic positions *j*, f_j *is* the scattering factor of the *j*th atom and x_j, y_j, z_j are the positional coordinates of the *j*th atom (Rhodes, 1993).

The structure factor that describes the reflection *hkl* is a composite wave created by the superposition of many individual waves, each resulting from diffraction by an individual atom, specifically by its electron clouds, which means that the diffraction reveals the distribution of electrons or electron density of the molecules and describes the surface features and overall shape of all the molecules in the crystal (Rhodes, 1993). Because protein molecules are ordered in a crystal, the electron density can be described mathematically by a complex three-dimensional periodic function also described as a Fourier series:

$$\boldsymbol{\rho}_{xyz} = \frac{1}{V} \sum_h \sum_k \sum_l |F_{hkl}| \, e^{[-2\pi i(hx + ky + lz) + i\alpha_{hkl}]}$$

where $\boldsymbol{\rho}_{xyz}$ is the electron density at every position *x,y*, *z* and describes the contents of the unit cells averaged over the whole crystal; V is the volume of the unit cell, $|F_{hkl}|$ is the structure factor amplitude of the reflection *hkl,* and α_{hkl} is the phase angle.

The view of the electron density as the Fourier transform of the structure factors implies that measuring the amplitude and the phase of each reflection would lead to obtaining the electron density that could then be displayed by a computer (Rhodes, 1993). Unfortunately, during a diffraction experiment, while the intensities (I) of the waves diffracted from the crystal are directly measured and give the amplitude of the structure factors ($I_{hkl} = |F_{hkl}|^2$), no information is available on the phase angles, as they cannot be directly recorded so are lost during the experiment. This is known as 'The Phase Problem' and represents the major obstacle to constructing an initial structural model (Taylor, 2003).

Initial phase estimates can be obtained in a variety of ways such as: Direct methods (*ab initio*) usually used for small molecules only as they require very high resolution data (<1.2Å); Molecular replacement for which the similarity of the unknown structure to an already known structure is a prerequisite; (Multiple or Single) Isomorphous Replacement (MIR or SIR), which requires the attachment of heavy atoms to the

protein molecules in the crystal (by soaking or co-crystallisation) followed by analysis of the differences between the data with and without the heavy atoms; and finally, the (multi or single wavelength) Anomalous Diffraction method (MAD or SAD) which depends on the presence of sufficiently strong anomalously scattering atoms in the protein structure itself and a X-ray radiation wavelength at or near the absorption edge of that element (Drenth, 1994; Taylor, 2003).

In recent years, improved X-ray sources, detectors and software have led to the routine use of anomalous scattering to obtain phase information (Taylor, 2010). Selenium is the most commonly used anomalous scatterer for MAD or SAD studies as the X-ray absorption edge of selenium (~0.98 Å) is accessible with the tunable wavelength radiation produced in synchrotrons; also, it can be easily added (in the form of selenomethionine) into proteins over-expressed in *E. coli* grown in medium containing selenomethionine as the only source of methionine (Doublié, 2007).

In the best cases, only an accurate single set of X-ray data (SAD) collected at just one wavelength (typically at the absorption peak) is required to provide the positions of the anomalous scatters. However, it is not formally possible to evaluate a protein phase exactly if there are only two experimental measurements, (*i.e.* with only a single anomalous difference available in SAD, or in the SIR case when only the native and one derivative data set is measured) due to the fact that there is a two-fold ambiguity in the estimation of the protein phase that needs to be broken (Dodson, 2003). The correct enantiomer can only be chosen by assessing which hand generates the better electron density map by means of density-modification procedures, which modify the initial density to generate a new set of phases that are then combined with the experimental ones successively; since the phases derived from the two enantiomorphs differ in quality, this procedure should also select the correct hand (Dodson, 2003; Rice *et al.*, 2000).

Then, with the measured structure factors and correct estimates of the phases (by any of the methods previously mentioned), an initial electron density map can be calculated, and interpreted in terms of amino acid structure with reference to the protein sequence. This initial structure is then subjected to an iterative model building, refinement and validation procedure that improves the agreement between the structural model and the experimental diffraction intensities. The model parameters that are optimised by a refinement program include, for each atom,

its *x*, *y* and *z* coordinates, and a parameter reflecting its 'mobility' known as the *B* factor, also referred to as 'temperature factor' (Wlodawer *et al.*, 2008).

Validation parameters that measure the global relative discrepancy between the experimentally obtained structure factor amplitudes, F_{obs}, and the calculated structure factor amplitudes, F_{calc}, obtained from the model are known as residuals, or *R*-factors. The R_{cryst} factor, defined as $\Sigma|F_{obs} - F_{calc}|/\Sigma F_{obs}$, gives an indication of the deviation of the model from reality . R_{free} is calculated analogously to the normal R_{cryst}, but for only 5-10% of the dataset, randomly selected, reflections which have never entered into model refinement, although they might have influenced model definition. In this way, if the mathematical model of the structure becomes unreasonably complex, *i.e.* includes parameters for which there is no justification in the experimental data, R_{free} will not improve (even though the R_{cryst} may decrease), indicating over-interpretation of the data (Drenth, 1994; Wlodawer *et al.*, 2008).

In order to judge if a crystal structure is complete, a geometry validation programme such as MolProbity needs to be used as it can trace global deviations of stereochemical parameters from their expected values, providing a broad-spectrum solidly-based evaluation of the model quality (at both the global and local levels) with a 'MolProbity score' that reflects the crystallographic resolution at which those values would be expected (Chen *et al.*, 2010). A particularly useful tool given by MolProbity is the Ramachandran plot, showing the mapping of pairs of ϕ/ψ torsion angles of the polypeptide backbone against the expected contours, which have a strong validation power because their values are usually not restrained in the refinement. Lastly, the set of coordinates, describing the location of each non-hydrogen atom in the protein molecule are then deposited in the PDB (Wlodawer *et al.*, 2008).

4.1.2. β-Xylosidase-1 structure

Due to *G. thermoglucosidasius* TM242 β-xylosidase-1 being a completely novel structure with no homologues available in the PDB, it was not possible to use molecular replacement as a phasing method (nor the *ab initio* method as the protein is too large (727 aa), and the crystals diffracted to insufficient resolution). For this reason, several heavy atom derivatives were investigated instead: protein crystals were soaked with platinum chloride for different lengths of time, the protein was co-crystallised with a set of lanthanides, and selenomethionine-derivatised protein was also produced. The last method was the one that eventually led to determination of the structure.

A high-resolution (1.7 Å) crystal structure of the native apo-enzyme was solved via SAD phasing using the diffraction data from crystals of selenomethionine-labelled protein, and a lower resolution structure (2.6 Å) of the enzyme-substrate complex was solved by rigid body refinement of the unbound structure. These novel crystal structures are the first reported structures of a GH52 family member and revealed unpredicted similarity to other glycoside hydrolase folds.

This thesis chapter details the crystallisation and information obtained from the crystal structure of *G. thermoglucosidasius* TM242 β-xylosidase-1. The coordinates of both structures obtained have been deposited in the PDB (accession codes: 4C1O and 4C1P respectively) and an original paper has been published in *Acta Crystallographica Section D* (Appendix III).

4.2. MATERIALS AND METHODS

4.2.1. Protein expression and purification

β-Xylosidase-1 was expressed following the method previously described in 2.3.1.1 and purified following the methods previously described in 2.3.5.1.2 and 2.3.5.2. Pure protein (~95%) samples up to 20 mg/mL were achieved prior to crystallisation studies.

4.2.2. Pre-crystallisation test

In order to assess the protein suitability (sufficient concentration and purity of the sample), a pre-crystallisation test (PCT) was performed according to the manufacturer's instructions (Hampton Research, USA).

4.2.3. Crystallisation screens

Preliminary crystallisation experiments were carried out via the sitting-drop vapour-diffusion method using a Phoenix crystallisation robot (Art Robbins, UK) with the HT96 format crystallisation screens: JSCG+, Morpheus and Structure Screen I + II (MDL). For the crystallisation trials, three sitting drops of 300 nL were set for each screening condition, each containing different ratios of pure protein (~10 mg/mL protein in 50 mM Tris-HCl buffer pH 8.0 containing 150 mM NaCl,) to well solution equilibrated against 60 µL reservoir solution,. The plates were sealed with plastic film and incubated at 16°C for the length of the experiments and were visually inspected under a microscope every two days, and then weekly for several months.

Furthermore, the protein sample was also tested with the crystallisation screens: NeXtal Classics Suite, NeXtal Classics II Suite, NeXtal PEGs Suite, NeXtal PEGs II Suite, NeXtal PACT Suite, NeXtal $AmSO_4$ Suite, NeXtal PhClear Suite, NeXtal PhClear II Suite and NeXtal Protein Complex Suite (Qiagen GmbH, Hilden, Germany) at the High Throughput Crystallization Facility (HTX) of the EMBL-Hamburg, during the 8th Protein Expression, Purification and Characterisation, EMBO practical course (PEPC8).

4.2.4. Optimisation of crystallisation conditions

Crystal hits from the initial screens were reproduced and optimised manually on a larger scale via the hanging-drop vapour-diffusion method using 24-well XRL/Linbro crystallisation plates (MDL). Optimisation consisted of gradients of pH and precipitant and salt concentration based on the best initial hits.

For the crystallisation, drop sizes of 3 μL containing different ratios of pure protein to well solution were suspended and equilibrated against 1 mL reservoir solution. Each well was then covered with a plastic coverslip and sealed with high vacuum grease (Dow Corning, Cardiff, Wales).

The plate was incubated at 18°C for the length of the experiments and visually inspected regularly. Prior to the data collection, the crystals were carefully removed from the drops with a cryoloop using glycerol as cryoprotectant (10% (v/v) diluted in well solution), by Dr Susan Crennell or Dr Jean van den Elsen (University of Bath). Cryoloops containing crystals were then mounted onto the collection apparatus within a stream of nitrogen gas or flash-frozen in liquid nitrogen prior to transport to Deutsches Elektronen-Synchrotron, DESY, Hamburg, Germany.

4.2.4.1. Co-crystallisation with xylose

Co-crystallisation of β-xylosidase-1 with its product xylose was performed as described in 4.2.4 adding 100 mM xylose to the protein buffer. Sodium formate was used as a cryoprotectant (to a final concentration of 1.6 M in well solution containing 100 mM xylose).

4.2.5. Selenomethione derivatised protein

4.2.5.1. Recombinant expression of SeMet-labelled protein

E. coli BL21 harbouring the pET-28a(+)-β-xylosidase plasmid was grown overnight in 5 mL of LB medium containing kanamycin at 37°C with shaking at 220 rpm. Cells were collected by centrifugation, washed twice with sterile distilled water and then inoculated into 500 mL of sterile SelenoMethionine Medium Complete (MDL), in 2 L baffled flasks. Cells were grown at 37°C with shaking at 220 rpm until an OD_{600} of 1 was reached. The medium was then supplemented with lysine, phenylalanine and threonine at 100 mg/L, and isoleucine, leucine and valine at 50 mg/L, to block methionine biosynthesis in a non-auxotrophic *E. coli* strain (Doublié, 2007). Then, 2 mL of Selenomethionine 250x (10 mg/mL) were added 15 min prior to induction with 0.3 mM IPTG for 5 h. The cells were harvested, resuspended and lysed as described in 2.3.3, adding 1 mM tris-(2-carboxyethyl)phosphine (TCEP) (Calbiochem, Millipore Ltd) to the resuspension buffer to prevent Se-Met oxidation.

4.2.5.2. Mass spectrometry

Selenomethionine incorporation into the protein was verified through mass spectrometry by Mr Mervyn Lewis from the Department of Chemistry, University of

Bath. Native and SeMet-derivatised protein at 5 mg/mL concentration were analysed by a Nano LC-CHIP-MS system, consisting of the 1200 Nano HPLC-Chip microfluidic device (Agilent Technologies Inc., Santa Clara, CA, USA) coupled with the Agilent 6520 Accurate-Mass Quadrupole-Time-of-Flight Liquid Chromatography-Mass Spectrometry (Q-TOF LCMS). 1 µL of sample was loaded from the autosampler to the Agilent Chip enrichment column (5 µm ZORBAX 300SB-C3 (300 Å), 40 nL) at a loading flow of 4 µl/min, followed by passage to the analytical column (5 µm ZORBAX 300SB-C3 (300 Å), 75 µm × 43 mm) at 300 nL/min. The composition of the mobile phase A was 0.1% (v/v) formic acid and the liquid phase B consisted of 90% (v/v) acetonitrile and 0.1% (v/v) formic acid. Elution was obtained with increasing concentration of buffer B: 3% for the first 8 min, 50% for 2 min, and finally 100% for 2 min. Then the percentage of buffer B was decreased to 3% and maintained at this level to allow column re-equilibration. Samples were analysed by ESI in positive ion mode. Mass spectrometric data were acquired in the range m/z 100−3000 with an acquisition rate of 1.35 spectra per second, averaging 10,000 transients. The source parameters were adjusted as follows: drying gas temperature 300°C, drying gas flow rate 5 L/min, nebuliser pressure 45 psi, and fragmentor voltage 150 V. Data acquisition and processing were done using Agilent MassHunter Workstation Acquisition software (B.02.01 Build 2116).

4.2.5.3. SeMet-labelled protein crystallisation

As described in 4.2.4.

4.2.6. Enzyme assays

The enzyme was assayed spectrophotometrically using the chromogenic substrate *p*-nitrophenyl-β-D-xylopyranoside as described in 3.2.8, with the only difference being the enzyme buffer. SeMet-labelled protein was eluted from the size-exclusion column in buffer containing 50 mM Tris-HCl pH 8.0, 150 mM NaCl, 1 mM TCEP.

4.2.7. X-ray data collection and processing

4.2.7.1. Synchrotron radiation

Diffraction data were collected on EMBL beamline P13 at PETRA III, DESY by Dr Guillaume Pompidor. The beamline was equipped with a PILATUS 6M-F detector (DECTRIS, Baden, Switzerland) and an MD2 microdiffractometer (MAATEL, Moirans, France). Both SeMet-derivatised and native data sets were collected at λ= 0.9763 Å (12.700 keV), slightly above the Se K edge (12.658 keV), in shutterless data-collection

mode. For the SeMet-derivative crystals, a total of 540° was collected from a single crystal with an exposure time of 250 ms per 0.2° oscillation (25% beam transmission) at a crystal-to- detector distance of 343 mm. The native data set was measured over a total wedge of 140° with angular steps of 0.2°, an exposure time of 100 ms with a beam transmission of 100% and a crystal-to-detector distance of 319 mm. The data were processed using XDS (Kabsch, 2010) and were scaled with SCALA (Evans, 2006).

4.2.7.2. In-house X-ray source

Diffraction data from crystals of native enzyme co-crystallised with xylose were collected in-house at 100 K using a Rigaku X-ray generator MicroMax 007 HF with Saturn 944+ charge-coupled device detector at the wavelength corresponding to the Cu$K\alpha$ edge (1.5418 Å). The native data set was measured over a total of 288° with angular steps of 0.5°, an exposure time of 2 min per image and a crystal-to-detector distance of 95 mm. The data were processed with assistance of Dr Susan Crennell using the HKL 2000 package (Otwinowski & Minor, 1997).

4.2.8. Structure solution and refinement

The initial structure was obtained by Dr Thomas Schneider at EMBL via SAD-phasing and, after this initial model was obtained, the stages shown in Figure 4.1 were followed. The iterative manual model building was carried out in Coot (Emsley *et al.*, 2010) and the refinement of the model against the data was carried out using REFMAC5, part of the CCP4 suite, (Murshudov *et al.*, 2011) and PHENIX (Adams *et al.*, 2010) with final model evaluation using MolProbity (Chen *et al.*, 2010). All the above was performed with the assistance and expert supervision of Dr Susan Crennell, who carried out the final refinement stages.

The product-bound structure was solved by rigid body refinement of the apo-enzyme structure, and rebuilt and refined as above. Both structures have been deposited in the PDB and given the access codes 4C1O and 4C1P, respectively. Surface interface areas were determined using the PISA (Protein Interfaces, Surfaces, and Assemblies) server at the European Bioinformatics Institute (Krissinel & Henrick, 2007). PDB searching and structural comparison were performed using the DALI server (Holm & Rosenström, 2010) and PDBefold (Krissinel & Henrick, 2004). Images were created using the PyMOL Molecular Graphics System, Version 1.3 Schrödinger, LLC (Education-Use-Only PyMOL Builds). The solvent-accessible surface of a protein was calculated with AREAIMOL, part of the CCP4 suite.

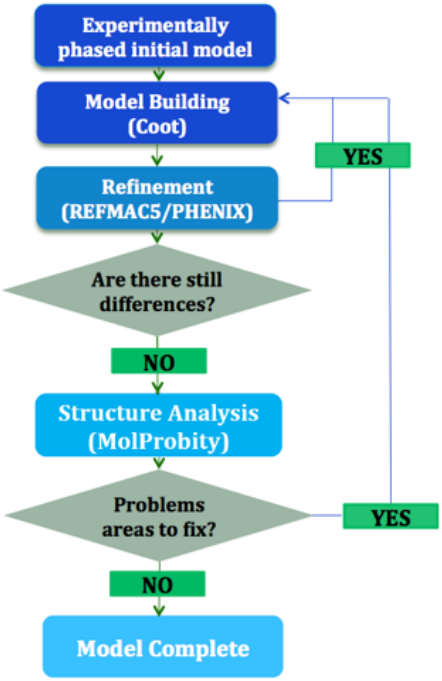


Figure 4.1. Structure solution and refinement flowchart. Scheme showing the different stages involved in obtaining a novel protein structure after an initial model has been obtained.

4.3. RESULTS AND DISCUSSION

4.3.1. Protein Crystallisation

The recombinant β-xylosidase-1 from *G. thermoglucosidasius* TM242 was expressed in a soluble, catalytically-active form in *E. coli*, and purified to homogeneity by nickel-affinity and gel filtration chromatography as shown in 3.3.3.1.1.

To find adequate conditions for the crystallisation of β-xylosidase-1, several high throughput crystallisation screens (3 in-house and 9 at the HTX of the EMBL-Hamburg) were carried out with different concentrations of pure protein. This initial screening was successful as crystal hits (and granular precipitate) were observed in every plate, with thin needles and needle clusters (hedgehog shaped crystals) appearing within 1-3 weeks. Representative drops are shown in Figure 4.2.

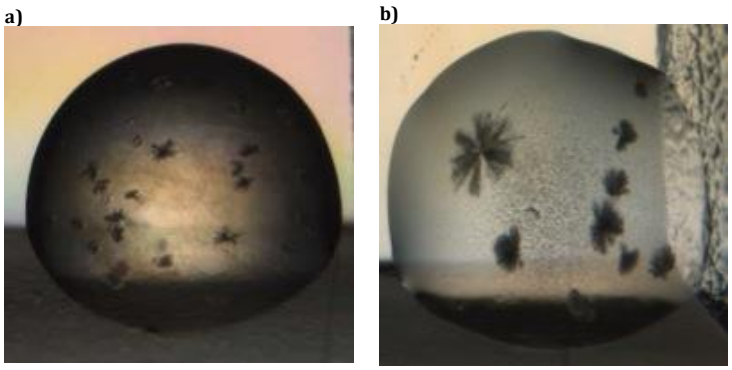


Figure 4.2. Examples of crystals obtained from high throughput crystallisation screening: 300 nl drops: (a) From well F7 Nextal Classics II Suite (containing 0.2 M ammonium sulphate, 0.1 M Bis-Tris pH 6.5 and 25% (w/v) PEG 3350), and (b) From well C10 Nextal Protein Complex Suite (containing 0.15 M ammonium sulphate, 0.1 M MES pH 5.5 and 25% (w/v) PEG 4K.

The best crystallisation conditions fell in the range pH 5.5 to 7.0, using different percentages of PEG (2K- 8K) as precipitant, with some needing concentrations between 0.15 to 1 M of salts (generally ammonium or lithium sulphate). These conditions were then varied systematically using different protein concentrations, to improve the size and quality of the crystals. After several optimisation attempts, plate-shaped crystals, mostly intergrown (Figure 4.3), were obtained after 3 weeks at 18°C.

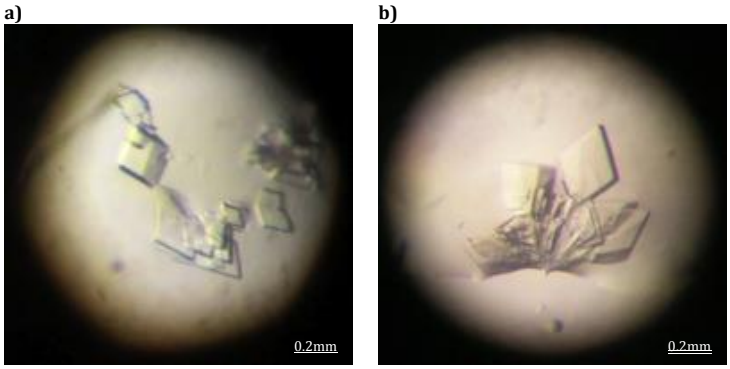


Figure 4.3. Examples of crystals obtained after optimisation of the crystallisation conditions: (a) Crystal from well containing 0.2 M Ammonium sulphate, 0.1 M MES pH 5.8 and 30% (w/v) PEG 4K, and (b) Crystal from well containing 0.4 M Ammonium sulphate, 0.1 M MES pH 6.0 and 25% (w/v) PEG 3350.

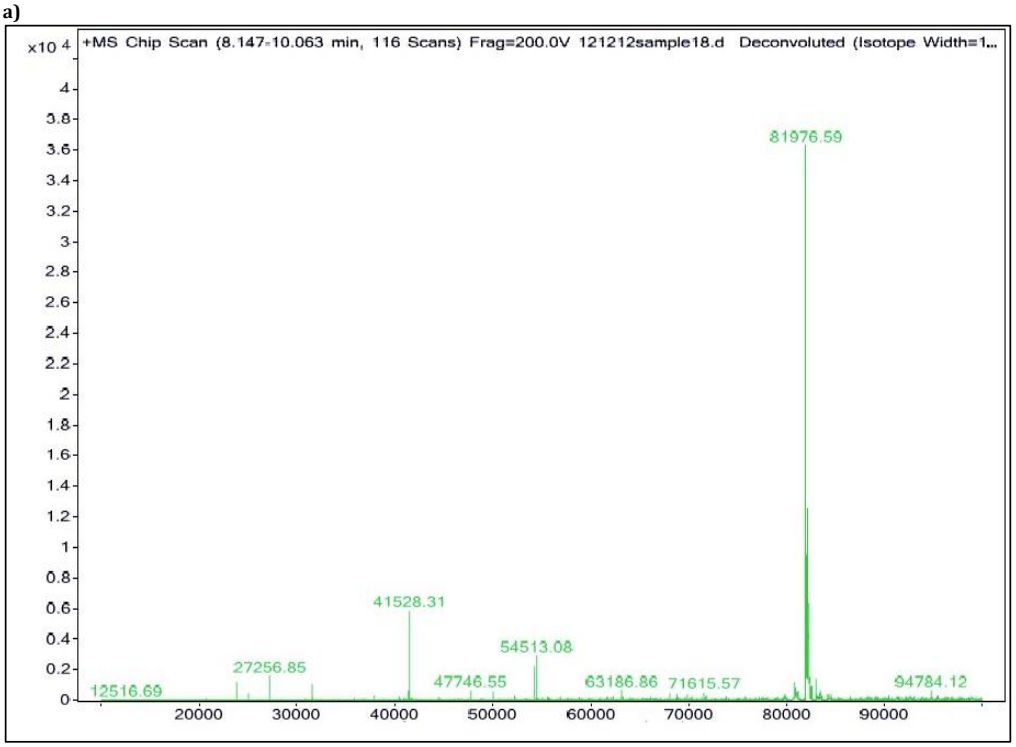
4.3.2. Protein derivatives

Due to this protein being a novel structure with no model available for molecular replacement, it was necessary to find other ways to solve the phase problem. As mentioned in the Introduction to this Chapter, this problem can also be solved by experimental phasing, using anomalous diffraction or the isomorphous replacement method.

β-Xylosidase-1 was recombinantly expressed by *E. coli* grown in minimal medium containing selenomethionine as the only source of methionine, purified and analysed by mass spectroscopy to check the degree of selenomethionine incorporation. The protein sequence of β-xylosidase-1 revealed 20 methionine residues (2.8%). Mass spectrometric analysis of the non-labelled recombinant protein gave a molecular mass (M_r) of 81,977 (Figure 4.4a) suggesting that the N-terminal methionine had been removed (expected M_r minus methionine = 81,975), while an M_r of 82,867 (Figure 4.4b) was determined for the SeMet derivatised protein, indicating that the 19 methionine residues available had been replaced by selenomethionine (expected M_r = 82,868).

The SeMet-derivatised enzyme was also catalytically active with a specific activity of 38 U/mg using *p*NP-β-D-xylopyranoside as substrate (comparable to 42 U/mg obtained with the non-labelled protein under the same assay conditions). The crystallisation conditions used were the same for labelled and non-labelled protein; that is, crystals were obtained by mixing β-xylosidase-1 (10 mg/mL in 50 mM Tris-HCl buffer, pH 8.0, 150 mM NaCl) with an equal volume of the well solution of 0.1 M MES buffer pH 6.0, 4 M ammonium sulphate and 25% (w/v) PEG 3350. Large crystals of approximately 0.1 x 0.05 x 0.05 mm (Figure 4.3) appeared after 3 weeks at 18°C.

a)



b)

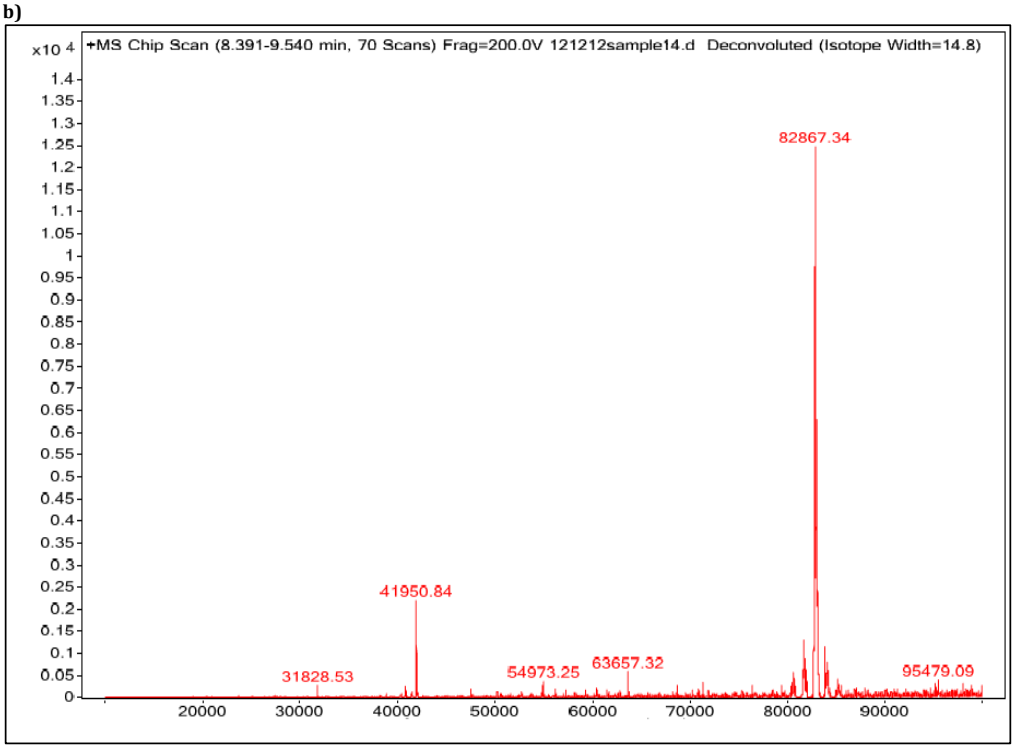


Figure 4.4. Analysis of SeMet incorporation by mass spectrometry (Nano LC-CHIP-MS system). (a) Mass spectrum of pure β-xylosidase recombinantly expressed in Terrific Broth medium. (b) Mass spectrum of pure β-xylosidase recombinantly expressed in M9 minimal medium supplemented with selenomethionine.

Considering that anomalous diffraction with SeMet-derivatised protein does not always work for phase determination (Boggon & Shapiro, 2000), other phasing attempts were performed in parallel: heavy atom derivatisation for isomorphous replacement method was tried by soaking protein crystals with 10 mM potassium tetrachloroplatinate K_2PtCl_4 (which binds to methionines and histidines) for different lengths of time (2 min, 10 min, 6 h); and co-crystallisation with lanthanides (macrocyclic gadolinium, ytterbium and lutetium complexes) (Pompidor *et al.*, 2008), was carried out by Dr Guillaume Pompidor at EMBL.

4.3.3. Preliminary X-ray data collection of crystals.

Preliminary diffraction data were collected and processed to 2.7 Å on beamline P14 of PETRA III synchrotron at DESY during the EMBO PEPC8 course, indicating the native crystals were orthorhombic and belong to the space group $C222_1$ with cell dimensions a=73.20 Å, b=104.92 Å, c=196.52 Å, α= 90°, β= 90°, γ= 90°.

These results were very different from the ones obtained by Czjzek *et al.* (2004), when they described the crystallization and preliminary crystallographic analysis of a thermostable GH52 β-xylosidase, XynB2, from *G. stearothermophilus* T-6, which shares 86% identity with the β-xylosidase-1 from TM242. Their native crystals were triclinic, diffracted to 2.0 Å and the unit-cell parameters were a= 80.6 Å, b= 97.5 Å, c=107.2 Å, α= 107.4°, β= 98.2°, γ= 106.6° in space group P1 (Czjzek *et al.*, 2004). As no structure for this XynB2 enzyme has been reported during the past 10 years, difficulties in obtaining the phases might be speculated.

4.3.4. X-ray data collection of SeMet protein crystals.

In order to obtain the phases, SeMet-derivatised protein crystals (flash frozen using 10% (v/v) glycerol as cryoprotectant, mixed with the crystallisation well solution of 0.1 M MES buffer pH 6.0, 4 M ammonium sulphate and 25% (w/v) PEG 3350) were exposed to X-rays on beamline P13 of PETRA III synchrotron. Statistics of the data collection are shown in Table 4.1

Table 4.1 Data-collection statistics from a SeMet derivatised protein crystal (data collected at λ= 0.9763 Å)**:**

Data collection statistics	Glycerol
Space group	$C222_1$
Unit-cell parameters (Å, °)	a=73.56, b=105.10, c=195.60 α = 90 β = 90, γ = 90
Resolution (Å)	1.7(1.73-1.70)
Total No. of reflections	429091
No. of unique reflections	83431
Completeness (%)	99.9 (100.0)
Redundancy	5.1(5.3)
<I/σ(I)>	18.6 (2.7)
R_{merge}	0.045 (0.530)
Overall average *B* factor from Wilson plot (Å^2)	19.6

where <I/σ(I)> is the average signal-to-noise ratio and $R_{merge}= \Sigma_{hkl}\Sigma_j |I_{hkl,j} - <I_{hkl}>| / \Sigma_{hkl}\Sigma_j |I_{hkl,j}|$ where $I_{hkl,j}$ is the *j*th observation of reflection *hkl* and $<I_{hkl,j}>$ is the mean intensity for all observations of I_{hkl}.

4.3.5. Phasing, model building and refinement

The structure was solved via SAD phasing using the SeMet anomalous diffraction data in β-test versions of SHELXC/D/E (Sheldrick, 2008) via a β-test version of HKL2MAP (Pape & Schneider, 2004) by Dr Thomas Schneider at EMBL. For solution of the selenium substructure, 20 sites were requested and the data were truncated to 3.0 Å resolution; for density modification a solvent content of 46% was specified, but for all other parameters default values were used. The electron density map resulting from SHELXE after refinement of the selenium sites and ten macro-cycles consisting of 20 cycles of density modification followed by auto-tracing was automatically interpreted by ARP/wARP (Langer *et al.*, 2008) building 697 residues (out of 727). Solving the β-xylosidase structure by using this method meant that the parallel experiments with heavy atoms soaking and lanthanides co-crystallisation were no longer needed and discontinued.

The selenomethionine data were subjected to rigid body refinement in REFMAC5 and the reliability factors were R_{cryst}= 0.2461, R_{free}= 0.2473. Manual model building was done with Coot (Emsley *et al.*, 2010) under the supervision and assistance of Dr Susan Crennell. Several different non-protein molecules present in the crystallisation condition (seventeen glycerols, six PEGs, four sulphates, five 1,2-ethanediols, one sodium) and 637 waters were seen in the electron density, and dual conformations were observed for 54 residues in the final structure. The model was subjected to several rounds of refinement and validation using REFMAC5 and PHENIX, followed by rebuilding where necessary; this process continued until the structure was judged to be

completed. Final model evaluation was done using MolProbity (Chen *et al.*, 2010). The final structure statistics are shown in Table 4.2.

Table 4.2. β-Xylosidase structure refinement statistic:

Structure Refinement	**Glycerol**
Resolution range (Å)	97.78-1.70
No. of reflections, working set	83366
No. of reflections, test set	4165
Final R_{cryst}	0.1404
Final R_{free}	0.1730
Number of non-H atoms	
Protein	5862
Ligands/ions	167/21
Water	637
R.m.s deviations	
Bonds (Å)	0.010
Angles (°)	1.261
***B* factors (Å²)**	
Protein	23.55
Ligand/ions	47.56/69.10
Water	35.59
Ramachandran plot analysis	
Most favoured regions	97.5 %
Disallowed regions	0
Molprobity score	1.12 99th percentile* (N=9248, 1.70Å ± 0.25Å)
Clashscore	2.64 99th percentile* (N=819, 1.70Å ± 0.25Å)

At this point of the refinement process, the reliability factors were R_{cryst}= 0.1404, R_{free}= 0.1730 and the Molprobity score =1.12; the Ramachandran analysis (shown in Figure 4.5) indicates that 97.3% of all residues were in favoured regions, 100% of all residues were in allowed regions, and there were no outliers (Lovell *et al.*, 2003). The coordinates were deposited in the PDB with the accession code 4C1O and, according to the wwwPDB X-ray Structure Validation Report, there were no chirality outliers and the only planarity outlier was Tyr712.

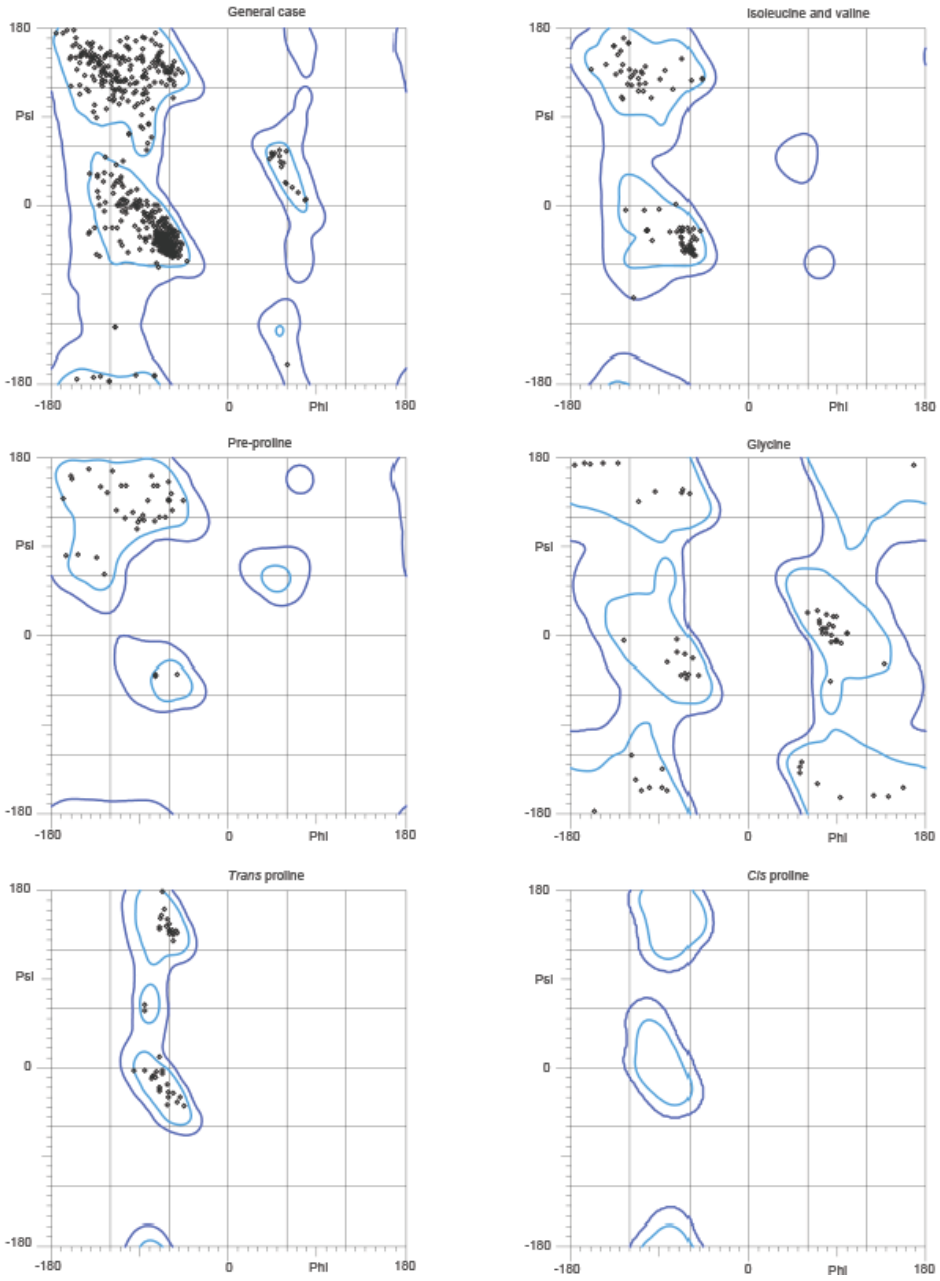


Figure 4.5. MolProbity Ramachandran analysis of β-xylosidase structure. The plots shows that 97.3% (760/781) of all residues are in favoured (98%) regions and 100% (781/781) of all residues are in allowed (>99.8) regions. There are no outliers. Data obtained from kinemage.biochem.duke.edu (Lovell *et al.*, 2003).

4.3.6. β-xylosidase structure

The structure of β-xylosidase reveals that the protein has two domains, an N-terminal domain folding into a β-sandwich and a C-terminal $(\alpha/\alpha)_6$ barrel.

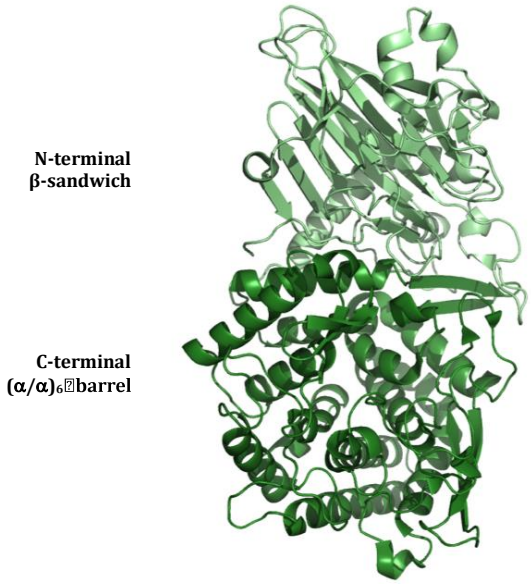


Figure 4.6. The first GH52 crystal structure. β-xylosidase-1 has two domains, an N-terminal domain folding into a β-sandwich and a C-terminal $(\alpha/\alpha)_6$ barrel.

Surface assemblies were determined using PISAserver (Protein Interfaces, Surfaces, and Assemblies) at the European Bioinformatics Institute (EBI) (Krissinel & Henrick, 2007) and the result is given in Table 4.3. Other interfaces had less than 1,000 Å^2 so were not judged significant.

Table 4.3. Probable assemblies in β-xylosidase-1 determined by PISA server.

PQS set	mm Size	Formula and Composition	Stable	Surface area [Å^2]	Buried area [Å^2]	ΔG^{int} [kcal/mol]	ΔG^{diss} [kcal/mol]
		$A_2a_8b_2c_{10}d_{34}e_{10}f_2$					
1	2	$A_2[SO4]_8[NA]_2[EDO]_{10}[GOL]_{34}[PEG]_{10}[PGE]_2$	Yes	46340	18560	-89.2	18.4

PQS set= Probable Quaternary Structure, group of different assemblies that possibly make the crystal.
mmSize= number of macromolecular monomeric units in the corresponding assembly, refered to as the oligomeric state.
Formula= indicates the chemical composition of the assembly.
Composition= indicates monomeric units found in the assembly.
Surface area= indicates the solvent-accessible surface area of the assembly in square Å.
Buried area= indicates the solvent-accessible surface area of monomeric units buried upon assembly formation.
ΔG^{int}= indicates the solvation free energy gain upon formation of the assembly.
ΔG^{diss}= indicates the change in free energy of assembly dissociation.

The analysis indicated that the β-xylosidase-1 exists as a stable dimer within the crystal, the symmetry-related molecule being rotated so the N-terminal domain of one molecule interacts with the C-terminal domain of the other (Figure 4.7), burying 18,560 $Å^2$ of surface, with ten hydrogen bonds, four salt bridges and a significant contribution of hydrophobic interactions across the interface.

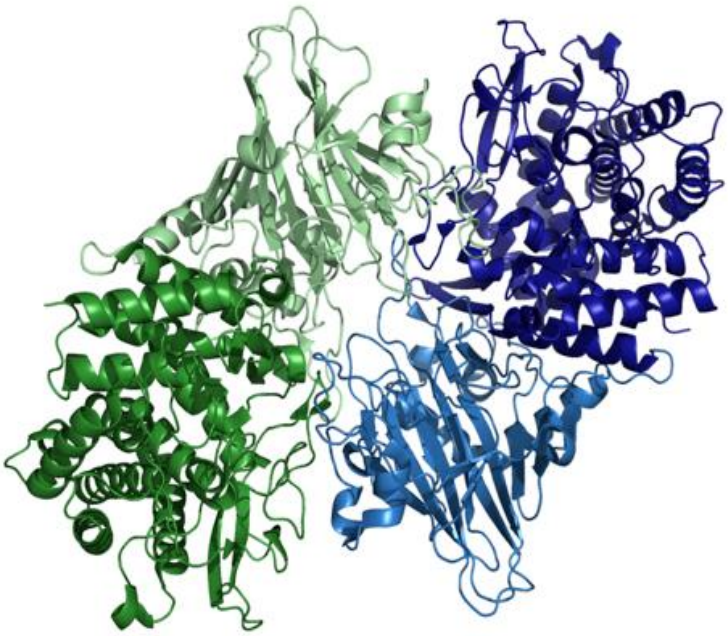


Figure 4.7. A cartoon representation of the β-xylosidase dimer. Each N-terminal domain is shown in lighter colours and each C-terminal domain in darker colours.

The oligomeric state of the protein is consistent with preliminary analysis by SAXS as part of the EMBO practical course. The SAXS data were collected on EMBL beamline X33 at DORIS III (DESY) and the data were processed and analysed by Dr Dmitri Svergun who concluded that the β-xylosidase enzyme is a dimer in solution (personal communication). This is also supported by the results obtained from size-exclusion chromatography (SEC) shown in the previous chapter (Section 3.3.3.1.1.), where comparison of the elution of the β-xylosidase with that of standard proteins of known molecular masses gave an M_r of 168 kDa for β-xylosidase, corroborating the dimerisation of the protein.

Furthermore, all the β-xylosidases belonging to GH52 characterised to date in CAZy: XylA from *Aeromonas caviae* ME-1 (Suzuki *et al.*, 2001), XynB2 from *Geobacillus stearothermophilus* 21 (Nanmori *et al.*, 1990), XynB2 from *G. stearothermophilus* T-6 (Bravman *et al.*, 2001), XylA from *Paenibacillus* sp. DG-22 (Lee *et al.*, 2007), have been

found to be dimeric (mostly through SEC). Contreras *et al.* (2008) published the first extensive conformational characterisation of XynB2 from *Geobacillus stearothermophilus,* in which, along with the results obtained from SEC, Isothermal Titration Calorimetry (ITC) and analytical ultracentrifugation were performed to define the oligomeric state of the protein. They concluded that the enzyme is in fact a dimer, and further analysis showed that this assembly is essential for efficient catalytic activity as the active site of the enzyme is formed by the two monomers. They also found that probably aromatic residues are involved in the catalysis (Contreras *et al.*, 2008).

4.3.6.1. β-xylosidase putative active site

Interestingly, a glycerol molecule (from the cryoprotectant solution) was found to be tightly bound in a cleft in the $(\alpha/\alpha)_6$ domain and on the interface between the two monomers (Figure 4.8). The average *B* factor (indicative of the mobility of the individual atoms) for the tightly-bound glycerol was less than half of the mean *B* factor for the 17 glycerols bound in the structure; 22.4 Å^2 compared with 48.3 Å^2). A second glycerol molecule is bound outside the cleft but close to the first tightly-bound glycerol (average *B* factor of 32.2 Å^2) with hydrogen bonds being formed across the dimer interface.

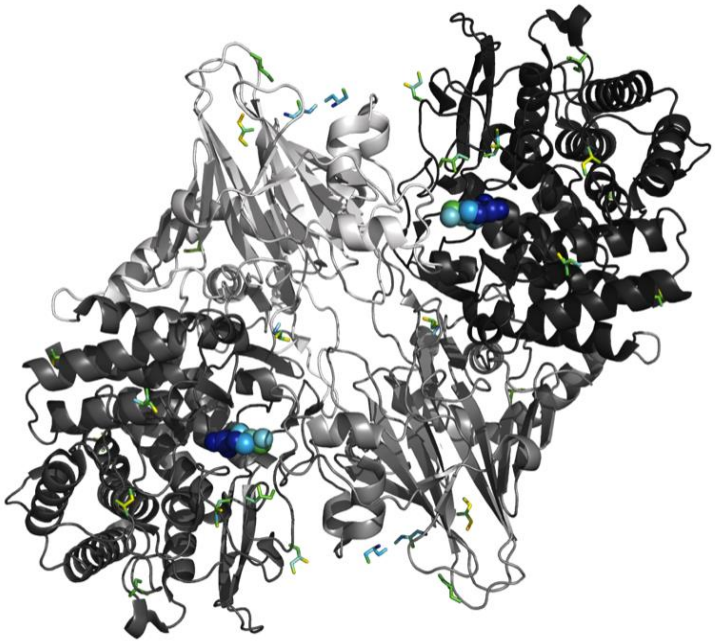


Figure 4.8. A cartoon representation of the β-xylosidase dimer with glycerols bound. Each N-terminal domain is shown in lighter colours and each C-terminal domain in darker colours. The glycerol molecules bound in the structure are shown in stick form with the two tightly-bound in the putative active sites shown in space-filling representation coloured according to their *B* factors, from dark blue for low *B* factor to red for high *B* factor.

Glycerol can take up a conformation mimicking half of a sugar ring, and therefore these two tightly-bound glycerols may be binding at the active site. Kinetic data (Table 4.4 and Figure 4.9) supports this; that is, in the presence of glycerol the observed K_M of the enzyme increases (almost double), while the *V*max remains practically the same, characteristic of a competitive inhibitor that binds to the active site of the enzyme and prevents binding of the substrate.

Table 4.4. Kinetic parameters of the recombinant β-xylosidase-1 with and without glycerol. Enzyme assays were carried out at 65°C. *V*max and K_M values were calculated using SigmaPlot 12 software. Standard Errors on all values are <10%.

	K_M (mM)	*V*max (µmol *p*NP/min/mg)
β-xylosidase (without glycerol)	0.37 ± 0.0148	42.1 ± 0.4041
β-xylosidase (with glycerol)	1.07 ± 0.0484	40.3 ± 0.4713

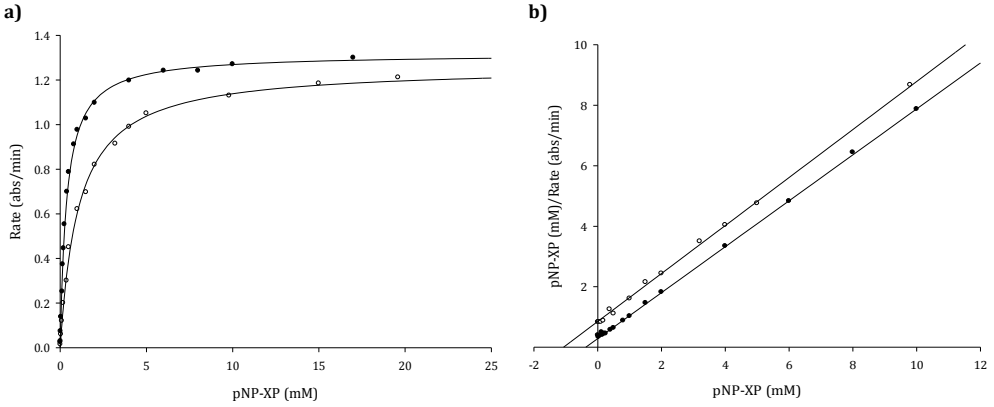


Figure 4.9. Effect of glycerol on β-xylosidase activity. The enzyme was assayed with different concentrations of the subtrate *p*NP-β-D-xylopyranoside in the absence (black circles) and presence (white circles) of glycerol. The enzymatic reaction velocity is given in A_{410}/min. (a) Michaelis-Menten plot of the reaction velocity (v) as a function of the substrate concentration [S]. (b) Hanes Woolf plot ([S]/v *versus* [S]), where the parallel lines suggest competitive inhibition.

The idea that the two tightly-bound glycerol molecules indicate the actual position of the enzyme active site is also supported by the observations that this cleft is on the interface between the two monomers, it contains aromatic residues and that two carboxylic amino acids, known to play a key role in the hydrolysis of the glycosidic bond by glycoside hydrolases, are positioned in between the two glycerol molecules, as can be seen in Figure 4.10.

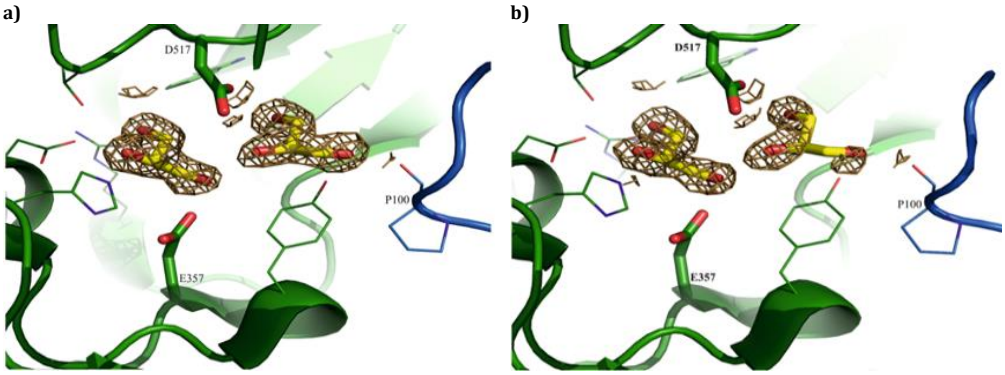


Figure 4.10. Glycerol bound in the active site of the β-xylosidase structure. The residues forming the cleft are shown in stick form and the carboxylic residues, E357 and D517, are shown in thicker lines with: (a) The $2F_{obs}$-F_{calc} electron density within 1.5 Å of the glycerol molecules contoured at 1.2σ. (b) The simulated-annealing F_{obs}-F_{calc} OMIT electron density within 1.5 Å of the glycerol molecules contoured at 1.2σ. The proline on the right-hand side of the active site comes from the partner subunit of the dimer.

4.3.7. Co-crystallisation with xylose

Subsequently, pure protein was co-crystallised with xylose under the same conditions previously described to confirm the position of the active site. Since the glycerol cryoprotectant is required at a concentration that would compete with substrate or product binding to the active site, formate was used as the cryoprotectant in the xylose co-crystal data collection.

4.3.8. X-ray data collection and processing of xylose co-crystals.

Diffraction data of the enzyme co-crystallised with xylose were collected in-house at 2.6 Å resolution. Crystals grown in the presence of xylose were orthorhombic and belong to the space group $C222_1$ with cell dimensions a=73.369 Å, b=105.335 Å, c=194.834 Å, α= 90°, β= 90°, γ= 90°. The statistics of the data collection are shown in Table 4.5.

Table 4.5. Data-collection statistics from a SeMet derivatised protein crystal.

Data collection statistics	Formate + Xylose
Space group	$C222_1$
Unit-cell parameters (Å, °)	a=73.369, b=105.335, c=194.834 α = 90 β = 90, γ = 90
Resolution (Å)	2.64(2.69-2.64)
Total No. of reflections	508871
No. of unique reflections	22164
Completeness (%)	97.3 (78.0)
Redundancy	5.0 (2.1)
$\langle I/\sigma(I)\rangle$	7.62 (1.73)
R_{merge}	0.138 (0.481)
Overall average *B* factor from Wilson plot (Å^2)	14.9

where $\langle I/\sigma(I)\rangle$ is the average signal-to-noise ratio and $R_{merge}= \Sigma_{hkl}\Sigma_j |I_{hkl,j} - \langle I_{hkl}\rangle| / \Sigma_{hkl}\Sigma_j |I_{hkl,j}|$ where $I_{hkl,j}$ is the *j*th observation of reflection *hkl* and $\langle I_{hkl,j}\rangle$ is the mean intensity for all observations of I_{hkl}.

The product-bound structure was solved by molecular replacement using the unbound structure as a model. The model was subjected to rigid body refinement in REFMAC5 and the reliability factors were R_{cryst}= 0.2729 and R_{free}= 0.2847. Manual model building was carried out in Coot (Emsley *et al.*, 2010), and several different non-protein molecules present in the crystallisation condition (one xylobiose, one PEG, four sulphates, five 1,2-ethanediols and two sodiums) and 306 waters were seen in the electron density; dual conformations were observed for 15 residues in the final structure. The model was subjected to several rounds of rebuilding, refinement and validation using REFMAC5 and PHENIX and this process continued until the structure was judged to be completed. Final model evaluation was carried out using MolProbity (Chen *et al.*, 2010). The final structure statistics are shown in Table 4.6.

Table 4.6. β-xylosidase structure refinement statistic.

Structure Refinement	**Formate + Xylose**
Resolution range (Å)	27.31-2.634
No. of reflections, working set	22136
No. of reflections, test set	1169
Final R_{cryst}	0.1823
Final R_{free}	0.2380
Number of non-H atoms	
Protein	5674
Ligands/ions	26/1
Water	306
R.m.s deviations	
Bonds (Å)	0.003
Angles (°)	0.805
B factors (Å^2)	
Protein	13.75
Ligand/ions	20.69/10.63
Water	12.5
Ramachandran plot analysis	
Most favoured regions	96.01%
Disallowed regions	1
Molprobity score	1.58 99th percentile* (N=6042, 2.63Å ± 0.25Å)
Clashscore	2.63 100th percentile* (N=226, 2.63Å ± 0.25Å)

At this point of the refinement process, the reliability factors were R_{cryst}= 0.1823, R_{free}= 0.2380 and the Molprobity score= 1.58; the Ramachandran analysis (Figure 4.11) indicates that 95.7% of all residues were in favoured regions and 99.9% of all residues were in allowed regions. The only Ramachandran outlier was the residue Pro594

(Lovell *et al.*, 2003). The coordinates were deposited in the PDB with the accession code 4C1P and, according to the wwwPDB X-ray Structure Validation Report, there were no chirality outliers nor planarity outliers.

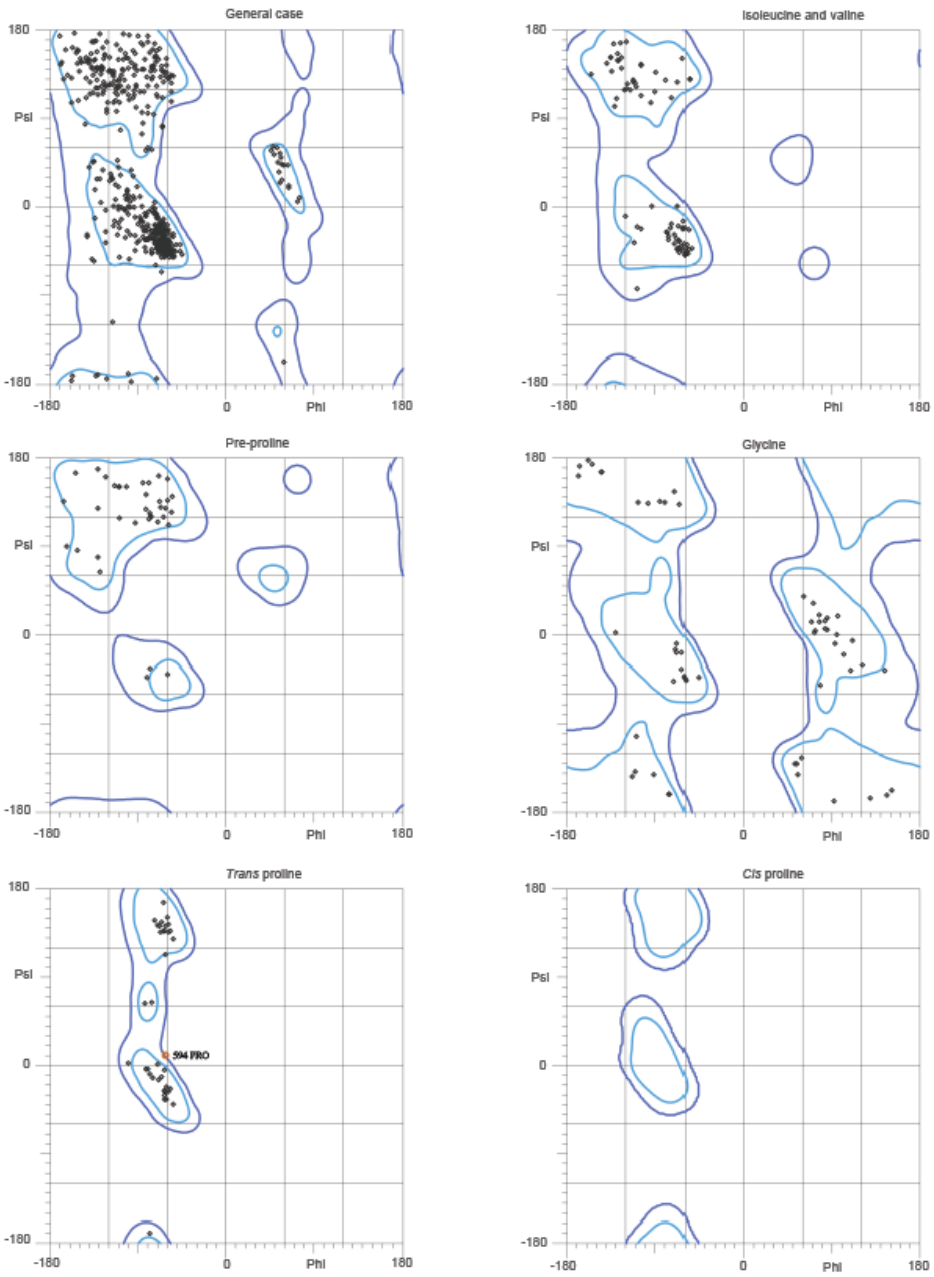


Figure 4.11. MolProbity Ramachandran analysis of β-xylosidase-xylose structure. The plots show that 95.7% (692/723) of all residues are in favoured (98%) regions and 99.9% (722/723) of all residues are in allowed (>99.8) regions. There is 1 outlier (ϕ/ψ): 594 Pro (-62.2, 10.7). Data obtained from kinemage.biochem.duke.edu (Lovell *et al.*, 2003).

4.3.9. The β-xylosidase catalytic site

The structure obtained from the crystals grown in the presence of xylose confirmed the cleft in the $(\alpha/\alpha)_6$ domain, where glycerol was previously found to be tightly bound, as the predominant xylose-binding site; xylose was not observed anywhere else in the structure. The difference Fourier map obtained revealed clear and sufficient electron density to fit two xylose molecules into the proposed active site; however, the continuity of the density was indicative of these two xylose units having been combined into a single molecule of xylobiose, as can be seen in Figure 4.12. This is entirely possible as xylose was in considerable excess over its K_M and presumably the enzyme has catalysed the reverse, glycosynthetic, reaction.

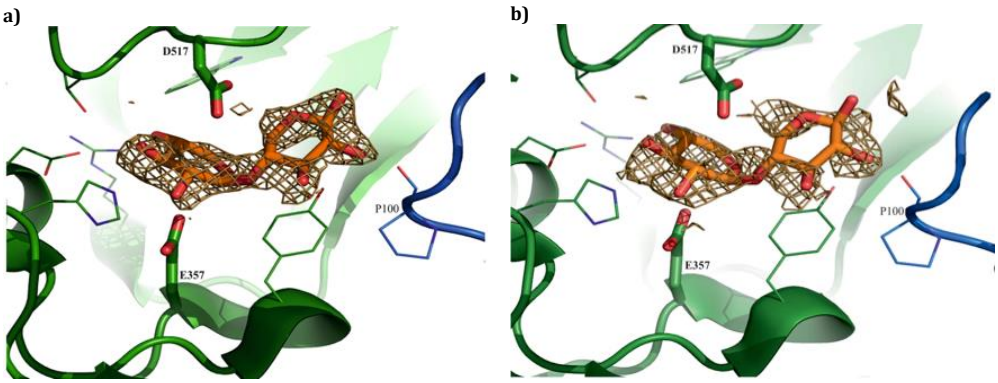


Figure 4.12. Xylobiose bound in the active site of the β-xylosidase structure. The residues forming the cleft are shown in stick form and the carboxylic residues, E357 and D517, are shown in thicker lines with: (a) The $2F_{obs}$-F_{calc} electron density within 1.5 Å of the glycerol molecules contoured at 1.2σ. (b) The simulated-annealing F_{obs}-F_{calc} OMIT electron density within 1.5 Å of the glycerol molecules contoured at 1.2σ. The proline on the right-hand side of the active site comes from the partner subunit of the dimer.

4.3.9.1. Mechanism of catalysis

β-xylosidases (and most other glycoside hydrolases) catalyse the hydrolysis of substrates by one of two canonical mechanisms originally outlined by Koshland (1953), inverting or retaining. Both reaction types employ a pair of carboxylic residues (Asp or Glu) in the active site, with one carboxylic acid group acting as a proton donor and the other as a nucleophile/base (Davies & Henrissat, 1995).

Retaining and inverting glycosidases differ in the distance separating their respective carboxyl groups, which determines whether hydrolysis of the glycosidic bond occurs with overall retention or inversion of the anomeric carbon configuration of the substrate. In both mechanisms the position of the proton donor is identical (being within hydrogen-bonding distance of the glycosidic oxygen), while the nucleophilic catalytic base is in close vicinity of the sugar anomeric carbon in retaining enzymes and

more distant in inverting enzymes, the latter accommodating a water molecule between the base and the sugar (Davies & Henrissat, 1995). This difference results in an average distance between the two catalytic residues of ~5.5 Å in retaining enzymes as opposed to ~10 Å in inverting enzymes (McCarter & Withers, 1994).

Retaining glycosidases catalyse the hydrolysis of the glycosidic bond by a two step double-displacement mechanism involving a covalent glycosyl-enzyme intermediate, with each step passing through an oxocarbenium ion-like transition state (Koshland, 1953; Zechel & Withers, 2001). During the first step (glycosylation), the acid/base catalyst acts as an acid and protonates the glycosidic oxygen concomitantly with bond cleavage and the nucleophile attacks at the anomeric center of the substrate sugar to form a covalent glycosyl-enzyme intermediate (Bravman *et al.*, 2001). In the second step (deglycosylation), the deprotonated acid/base catalytic group acts as a general base to activate a water molecule that carries out a nucleophilic attack on the glycosyl-enzyme intermediate, with the two inversion steps leading to retention of the stereochemistry at the anomeric center (Vuong & Wilson, 2010). In the case of retaining β-xylosidases, upon hydrolysis, the xylosyl glycone residue at substrate's non-reducing end is released as β-xylose, as shown in Figure 4.13 (Jordan & Wagschal, 2010).

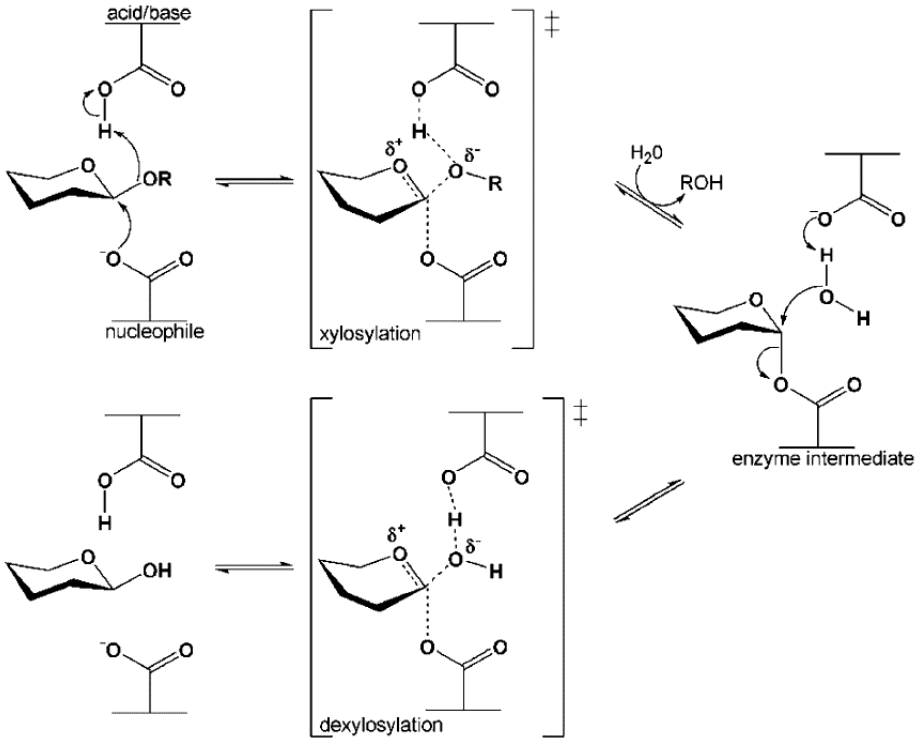


Figure 4.13. Retaining β-xylosidases (glycosidases) double-displacement reaction mechanism. Image obtained from Jordan & Wagschal(2010).

On the other hand, inverting glycosidases use a direct displacement mechanism to cleave the glycosidic bond in a single step, with one carboxylic acid group acting as a general base and activating a water molecule for attack at the anomeric carbon, while the other carboxylic residues acts as a general acid assisting the aglycone departure (Koshland, 1953; Zechel & Withers, 2001). In the case of inverting β-xylosidases, upon hydrolysis, the xylosyl glycone residue with a β-anomeric configuration at the substrate's non-reducing end is released as α-xylose, as shown in Figure 4.14 (Jordan & Wagschal, 2010).

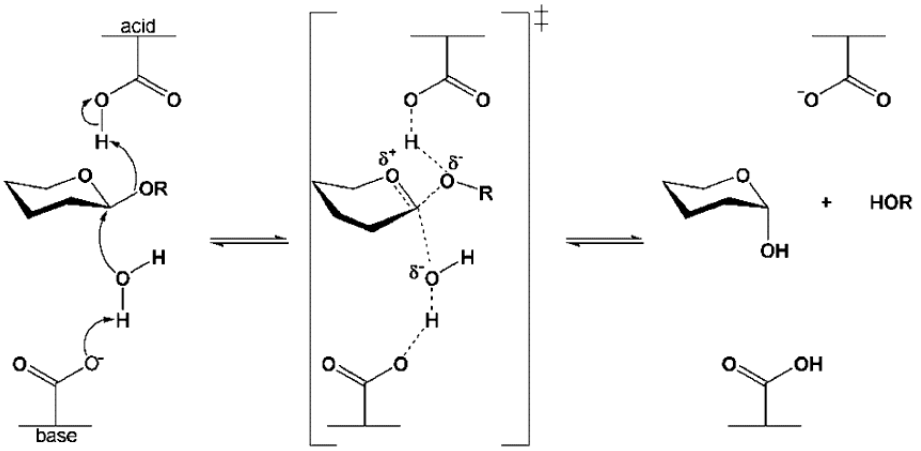


Figure 4.14. Inverting β-xylosidases (glycosidases) single-displacement reaction mechanism. Image from Jordan & Wagschal (2010).

Previous studies, in which the stereochemistry of the hydrolysis of *p*NP-XP by XynB2 (from *G. stearothermophilus* T-6) was followed by ^{1}H-NMR spectroscopy, showed that the configuration of the anomeric carbon was retained (Bravman *et al.*, 2001). This might indicate that β-xylosidases belonging to family GH52 possess a retaining mechanism of catalysis in which the hydrolysis of xylo-oligosaccharides to xylose occurs with the net retention of the anomeric carbon configuration, as enzymes within a family generally follow the same stereochemical course (with very few exceptions reported ((Vuong & Wilson, 2010)).

The crystal structure of the *G. thermoglucosidasius* TM242 β-xylosidase-1 reveals that, within the active site pocket, the two carboxylic residues positioned across the glycosidic bond between the xylose units, Glu357 and Asp517, are 6.5 Å apart (Figure 4.15), consistent with the distance expected for enzymes displaying a retaining mechanism of catalysis.

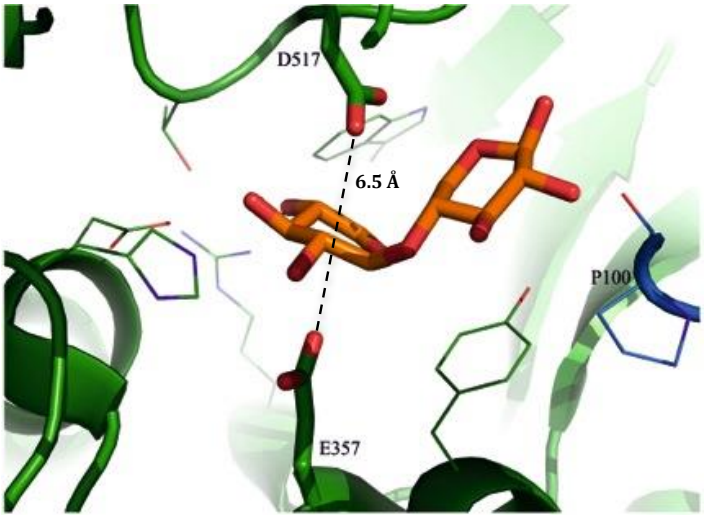
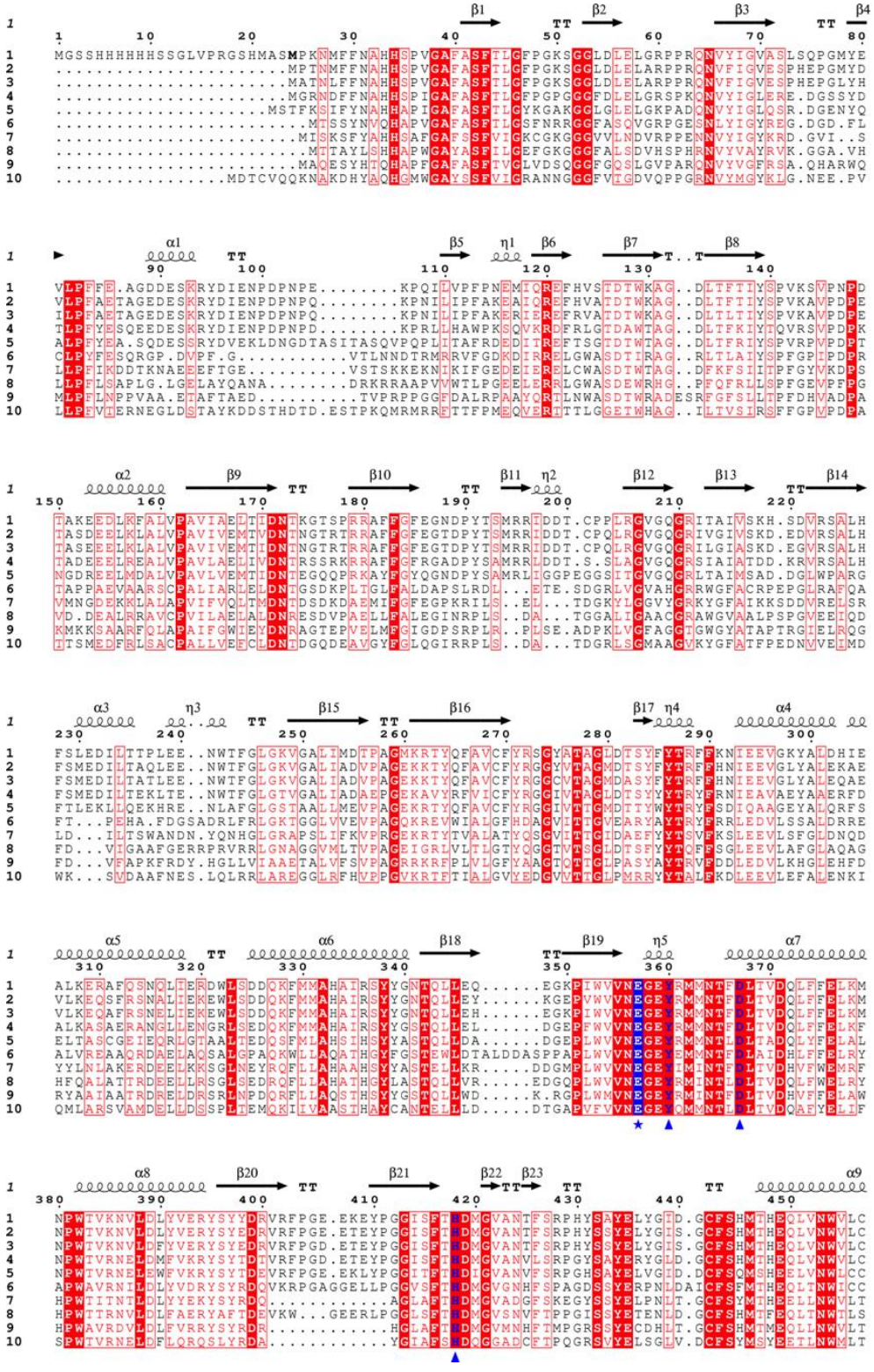


Figure 4.15. Xylobiose bound in the β-xylosidase active site. The residues forming the cleft are shown in stick form and the catalytic residues E357 and D517 are shown in thicker lines. The distance in between the carboxylic residues is shown as a dashed line measuring 6.5 Å.

Furthermore, in a sequence alignment of representative family 52 β-xylosidases (Figure 4.16), these two residues align perfectly with all members (highlighted in blue and indicated by blue stars) and correspond to Glu335 and Asp495 in XynB2 (sequence 3). Through site-directed mutagenesis of this enzyme, followed by detailed kinetic analyses including the pH-dependence of the catalytic constants, Bravman *et al.* (2003) have demonstrated that these residues are in fact the catalytic amino acids, with Glu335 acting as a nucleophile in the glycosyl-cleavage reaction and Asp495 assisting as an acid. Asp495 then acts as a base to catalyse the hydrolysis of the glycosyl-glutamyl intermediate.

As can be seen from Figure 4.16 not only the catalytic residues, but also the other amino acids in the active site environment, are conserved in all GH52 sequences available to date in CAZy (the one exception is XylA from *G. stearothermophilus* 236, but this sequence is 60 amino acids shorter than any other member of the family, making alignments difficult).



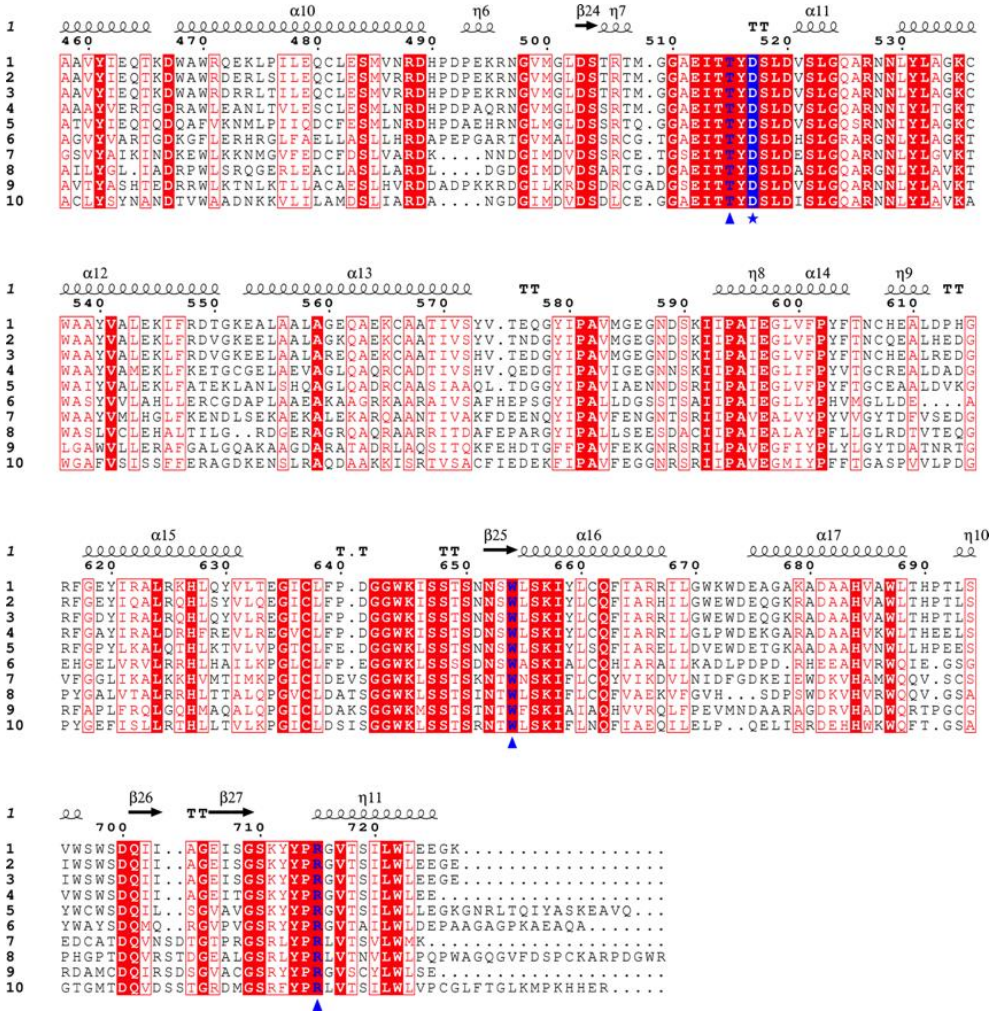


Figure 4.16. Structure-based sequence alignment of representative family 52 β-xylosidases available in CAZy, produced using Clustal Omega and displayed using ESPript 3.06 (Gouet *et al.*, 1999). Identical residues are shown in white on a red background, while conservative amino acid changes are shown in red. Amino acids involved in the interaction with substrate are highlighted in blue (the catalytic residues are indicated by blue stars, and the others by blue triangles beneath the alignment). Secondary structure of *Geobacillus thermoglucosidasius* TM242 β-xylosidase is shown above the alignment with β-strands as arrows and α-helices as coils. 1) *Geobacillus thermoglucosidasius* TM242; 2) *Geobacillus thermoglucosidasius* C56-YS93 (AEH48193.1, 87% identity); 3) *Geobacillus stearothermophilus* T-6 NCIMB 40222 (ABI49956.1, 86% identity); 4) *Paenibacillus* sp. DG-22 (ABV46494.1, 71% identity); 5) *Aeromonas caviae* ME-1 (BAA74507.1, 60% identity); 6) *Sorangium cellulosum* So0157-2 (AGP38099.1, 43% identity); 7) *Thermoanaerobacterium saccharolyticum* JW/SL-YS485 (AFK86458.1, 41% identity); 8) Deinococcus geothermalis DSM 11300 (ABF44123.1, 40% identity); 9) *Opitutus terrae* PB90-1 (AEC01950.1, 39% identity); 10) *Spirochaeta coccoides* DSM 17374 (AEC01950.1, 36% identity). It should be noted that the sequence of the *G. thermoglucosidasius* TM242 β-xylosidase is of the recombinant His-tagged protein, as crystallised. The native protein would begin at residue 24 (in bold) in this alignment.

4.3.9.2. Active site accessibility

From the crystal structure of the enzyme-substrate complex, it can be seen that there are two distinct xylose-binding subsites, one of which is buried and accommodates the -1 non-reducing end of xylose (Figure 4.17a), whilst the second is

more open, with its main interaction with the +1 reducing end of xylose being a non-directional stacking interaction with Tyr360 (Figure 4.17b). The -1 non-reducing end of xylose is held in an enclosed pocket by several hydrogen bonds (with most of the residues in the active site environment) and is distorted into a higher energy $^{4}H_{3}$ half-chair conformation, in marked contrast to the low- energy chair conformation taken up by the +1 xylose (Figure 4.17c).

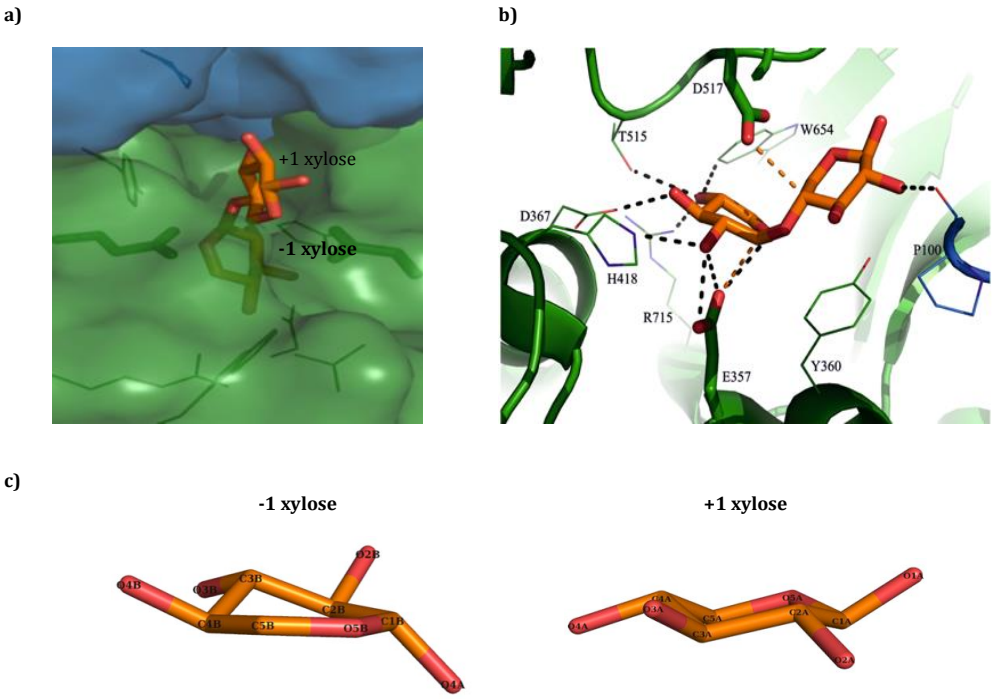


Figure 4.17. (a) A surface representation with 20% transparency of the β-xylosidase-1 active site cleft with the xylobiose substrate shown in stick form. The two subunits of the dimeric enzyme are shown in green and blue, the residues forming the cleft are shown in stick form, and the catalytic residues are shown in thicker lines. (b) Detail of the active site interactions, with hydrogen bonds shown as black dashed lines and carbon-oxygen interactions less than 3.2 Å shown in orange. (c) Details of the xylobiose forming molecules, their atoms and conformation; the -1 non-reducing end of xylose is shown in its higher energy $^{4}H_{3}$ half-chair conformation while +1 reducing end of xylose is shown in its the low- energy chair conformation.

A single hydrogen bond is made across the active site cleft between O2A (of the +1 xylose) and a main-chain atom from the N-terminus of the other monomer in the β-xylosidase-1 dimer, showing that the position of the active site at the interface restricts access to smaller xylo-oligosaccharides rather than bulk polymer xylan. This provides an explanation for the lack of catalytic activity of the enzyme towards xylan (shown in 3.3.3.1.2 and 5.3.4) and suggests that the catalytic efficiency of the enzyme might decrease when xylo-oligosacharides with a further degree of polymerisation are used as substrates. As seen in the previous chapter (3.3.3.1.2), the hydrolysis of

xylobiose by the β-xylosidase-1 is superior to xylotriose hydrolysis while activity against longer natural substrates like xylotetraose or xylopentaose is yet to be tested; however, considering the structural information, it can be speculated that these would be less effective substrates.

Recently, a GH52 β-xylosidase (GSxyn) from *Geobacillus stearothermophilus* 1A05585 has been successfully modified (by the mutation of Tyr509 to Glu509) to introduce a new catalytic xylanase activity, while conserving its β-xylosidase activity (Huang *et al.*, 2014). With no structure available at the time of the experiments performed by the authors, they did site-directed mutagenesis based on the amino acid sequence alignment of GSxyn and two other *G. stearothermophilus* xylanases where, in the same position, they have a glutamic acid instead of a tyrosine. GSxyn wild type displayed β-xylosidase activity using *p*NP-XP (K_M= 0.48 mM and k_{cat} = 36.6 s^{-1}), but no activity against beechwood xylan; on the other hand, Y509E mutant shows xylanolytic activity whilst retaining β-xylosidase activity (K_M= 0.51 mM and k_{cat} = 20.6 s^{-1}) (Huang *et al.*, 2014).

This Tyr509 residue corresponds to Tyr531 in the TM242 β-xylosidase-1 (which shares 86% sequence identity with GSxyn) and, according to the crystal structure, it does not belong to the active site (as it is 8.8 Å from the nearest OH on xylose) but it does make two strong H-bonds with Glu449. Huang *et al.* (2014) hypothesised that substituting this tyrosine with glutamic acid might introduce some flexibility into the orientation of the xylan chain in the substrate binding cleft and widen the active site pocket to allow xylan to access it. According to the structure, the introduced negative charge would repel Glu449 and thus disrupt the environment of the active site, providing a credible reason why the catalytic efficiency of the enzyme against *p*NP-XP decreased to almost half (from 76.3 $mM^{-1}\,s^{-1}$ in the wild-type to 40.5 in the mutant $mM^{-1}\,s^{-1}$; see individual K_M and k_{cat} values quoted above).

The introduction of broadened substrate acceptance into GSxyn has verified the possibility for engineering (through a single amino acid substitution) additional catalytic functions, sacrificing but without losing the original enzymatic activity (Huang *et al.*, 2014). As can be seen from Figure 4.16, most of the residues, including this tyrosine, are strongly conserved across the majority of GH52, suggesting members of this family have the potential to provide a structural scaffold for generating bifunctional enzymes.

4.3.10. Comparison with structures of enzymes with the same β-xylosidase function

β-Xylosidases, as identified by their EC number 3.2.1.37, are found in ten different glycosidases families in CAZy (Lombard *et al.*, 2014), as tabulated below (Table 4.7).

Table 4.7. Glycosidase hydrolase families classified in CAZy that contain at least one sequence with xylosidase (EC 3.2.1.37) activity: their mechanism of action, structure (with CATH classification where known), and a summary of the xylosidase structures deposited in the PDB, tbp refers to the structures that have not been published yet.

GH family	Mechanism	Domain(s) structure		Organism	β-xylosidase PDB entries
1	Retaining	$(\alpha/\beta)_8$ barrel 3.20.20.80			
3	Retaining	$(\alpha/\beta)_8$ barrel 3.20.20.300	Rossmann 3.40.50.1700		
30	Retaining	$(\alpha/\beta)_8$ barrel 3.20.20.80	β-sandwich 2.60.40.1180		
39	Retaining	$(\alpha/\beta)_8$ barrel 3.20.20.80	β-sandwich 2.60.40.1500	*Caulobactercrescentus* CB15	4EKJ (Santos *et al.*, 2012)
				Geobacillus stearothermophilus T-6 NCIMB 40222	1W91 (tbp) 2BFG & 2BS9 (Czjzek *et al.*, 2005).
				Thermoanaerobacterium saccharolyticum B6A-RI	1PX8 &1UHV (Yang *et al.*, 2004)
				Bacillus halodurans C-125	1YRZ (tbp)
				Bacillus subtilis 168	1YIF (tbp)
43	Inverting	β-propellor 2.115.10.20	β-sandwich 2.60.120.200	*Clostridium acetobutylicum* ATCC 824	1Y7B (tbp), 1YI7 (tbp), 3K1U (tbp)
				Geobacillus stearothermophilus T-6 NCIMB 40222	2EXH, 2EXI, 2EXJ, 2EXK (Brüx *et al.*, 2006)
				Selenomonas ruminantium GA192	3C2U (Brunzelle *et al.*, 2008)
51	Retaining	$(\alpha/\beta)_8$ barrel 3.20.20.80	Jelly Roll 2.60.40.1180		
52*	Retaining	$(\alpha/\alpha)_6$ barrel	β-sandwich		4C1O, 4C1P (Espina *et al.*, 2014)
54	Retaining	β-sandwich 2.60.120.200	β-trefoil 2.80.10.50		
116	Retaining	Unknown			
120*	Retaining	β-helix	β-sandwich	*Thermoanaerobacterium saccharolyticum* JW/SL-YS485	3VST, 3VSU, 3VSV (Huang *et al.*, 2012)

* these families contain only EC 3.2.1.37 xylosidase sequences

As can be seen from Table 4.7, the majority of these glycoside hydrolases families are of known fold (only GH116 remains with no known structure), although most of the structures deposited are of proteins with other glycosidase activities. β-Xylosidase structures are only available for members of families GH52 (this thesis), GH120, GH39, and GH43, all of which are shown in Figure 4.18.

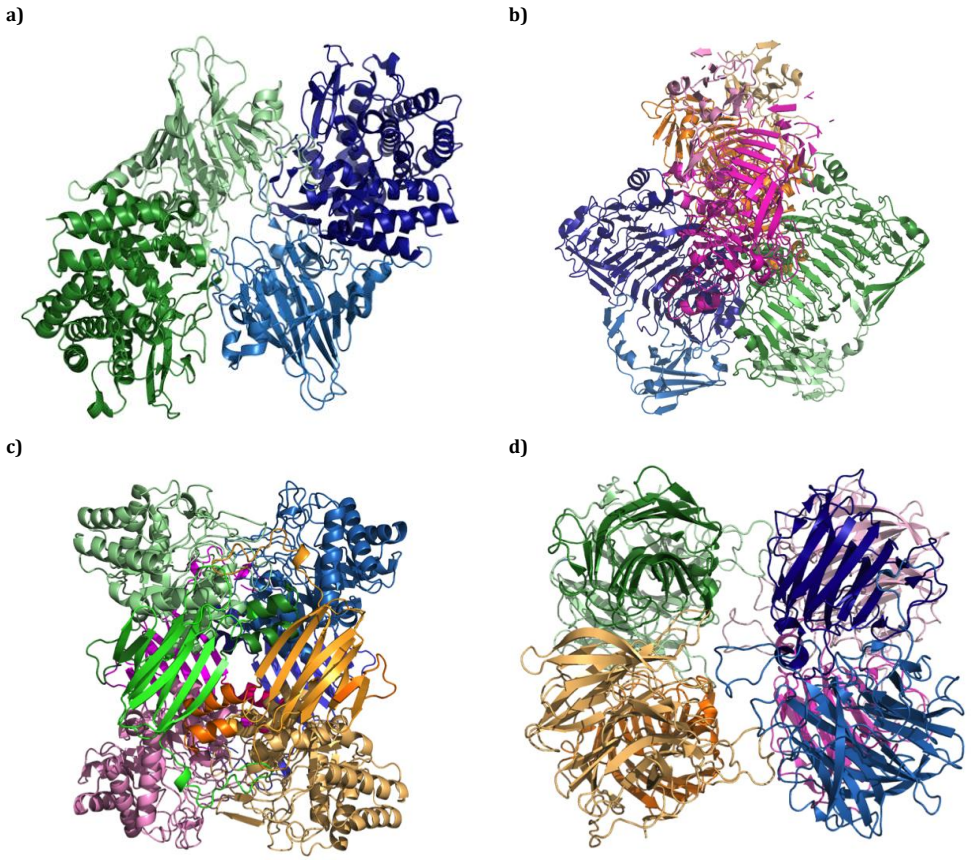


Figure 4.18. Cartoon representation of the different β-xylosidases protein folds. Each monomer is coloured differently, with the N-terminal domains shown in lighter colours and C-terminal domains shown in darker colours. (a) GH52 β-xylosidase-1: the protein structure is a dimer of two monomers aligned antiparallel to each other. Each monomeric subunit consists of an N terminal β-sandwich and a C-terminal $(\alpha/\alpha)_6$ barrel (Espina *et al.*, 2014). (b) GH120 XylC: the protein structure is a tetramer, and each monomeric subunit consists of a right-handed parallel β-helix core domain and a β-sandwich domain (Huang *et al.*, 2012). (c) GH39 XynB1: the protein structure is a tetramer and each monomeric subunit consists of a catalytic $(\alpha/\beta)_8$ barrel domain, a β-sandwich and a small C-terminal α-helical domain (Czjzek *et al.*, 2005). (d) GH43 XynB3: the protein structure is a tetramer made up of two dimers of two monomers aligned antiparallel to each other. Each monomeric subunit consist of an N-terminal β-propeller and a C-terminal β-sandwich (Brüx *et al.*, 2006).

The protein folds of these enzymes are completely different from that of the GH52 family described in Section 4.3.6; however, comparison of their active sites suggests common features between the GH52 and β-xylosidases belonging to GH39 and GH120, both families that hydrolyse with retention of configuration at the anomeric site. Family GH43 has not been included in the comparison since it contains inverting glycoside hydrolases, in which, as previously mentioned, the spatial positioning of the catalytic residues in the active site is totally different.

Detailed information about the interactions between each residue in the active site environment of the GH52, GH39 and GH120 β-xylosidases is listed in Table 4.8.

Table 4.8. Interactions shorter than 3.25 Å made by xylobiose in the xylosidase active site, and comparison with families 39 (data from Table 1 in Czjzek *et al.* 2005) and 120 (Huang *et al.*, 2012).

Xylobiose	GH52 BXP	Å	GH39 XYP/ANX	Å	GH120 BXP	Å
xylA stacking	Y360		Y283		H352	
O1A					Water	
O2A	O P100'	2.46				
O3A					$O^{\varepsilon 2}$ E405	3.17
C4	$O^{\delta 2}$ D517	3.13			$O^{\varepsilon 2}$ E405	2.59
O4A			$O^{\varepsilon 2}$ E160	2.84		
(O1 in ANX)			$O^{\varepsilon 1}$ E160	2.75		
C1B	$O^{\varepsilon 2}$ E357	2.99	$O^{\varepsilon 1}$ E278	3.22*	$O^{\delta 2}$ D382	3.10
O2B	$O^{\varepsilon 2}$ E357	3.11	$O^{\varepsilon 2}$ E160	2.90	$O^{\delta 1}$ D382	2.76
	$O^{\varepsilon 1}$ E357	3.14	$O^{\varepsilon 2}$ E278	2.40	$N^{\eta 1}$ R450	2.82
	$N^{\varepsilon 2}$ H418	2.99	$N^{\delta 2}$ N159	2.75		
			$O^{\delta 1}$ N159	2.88		
O3B	$O^{\delta 2}$ D367	2.89	$O^{\varepsilon 2}$ E324	2.62	$N^{\varepsilon 1}$ W383	2.93
			$N^{\varepsilon 2}$ H54	2.38	$O^{\varepsilon 1}$ Q289	2.62
					Water	2.95
O4B	$O^{\gamma 1}$ T515	3.00	$O^{\varepsilon 1}$ E324	2.53	$O^{\varepsilon 2}$ E353	2.57
	$N^{\eta 1}$ R715	3.00	$N^{\varepsilon 1}$ W316	3.18	$O^{\varepsilon 1}$ E353	3.09
	$C^{\zeta 3}$ W654	3.10			N^{ζ} K358	3.17
					$N^{\varepsilon 2}$ H360	2.83
					$O^{\delta 2}$ D382	3.18
O5B	$O^{\varepsilon 2}$ E357	3.12	O^{η} Y230	3.25	$O^{\delta 2}$ D382	3.00
					N^{ζ} K358	3.13
					O H352	3.16
xylB stacking					W113	

*Interaction visible in structure but not listed in Table 1 of Czjzek *et al.* 2005.

Comparison of *G. thermoglucosidasius* TM242 β-xylosidase-1 with a GH120 β-xylosidase which also has xylobiose bound (PDB entry: 3VSU (Huang *et al.*, 2012)) (Figure 4.19 a,b), reveals that both have a deep pocket binding the -1 xylose with many hydrogen bonds, while the +1 xylose is bound much less strongly, with few hydrogen bonds and stacking interactions with an aromatic residue. In both cases, this +1 site is in close proximity to another monomer of the oligomer, although interactions are only by a water-mediated hydrogen bond network in GH120. The closer proximity of the other monomer in *G. thermoglucosidasius* TM242 β-xylosidase-1 causes the active site cleft to be significantly narrower than in the GH120 structure. In both cases, the -1 sugar takes up a skew conformation, but the +1 xylose is in the relaxed chair conformation in β-xylosidase-1, while that in GH120 is in the higher energy boat form.

The active site of a GH39 β-xylosidase (PDB entry: 2BFG (Czjzek *et al.*, 2005)) seems to share most similarities with the *G. thermoglucosidasius* TM242 β-xylosidase-1 (Figure 4.19a,c); both have a deep cleft to accommodate the substrate, restricting access for larger substrates, although the cleft is within one monomer in GH39. If aligned by bound substrate, the nucleophiles (E278 and E357, respectively) also align, while the general bases do not. The buried -1 xylose is distorted into a similar conformation in both active sites, caused by a mixture of strong hydrogen bonds and hydrophobic

interactions and a stacking interaction in GH39. The conformation of the +1 xylose cannot be determined as *p*-nitrophenol occupies this site in the GH39 structure, but it is clear that both form stacking interactions with Tyr.

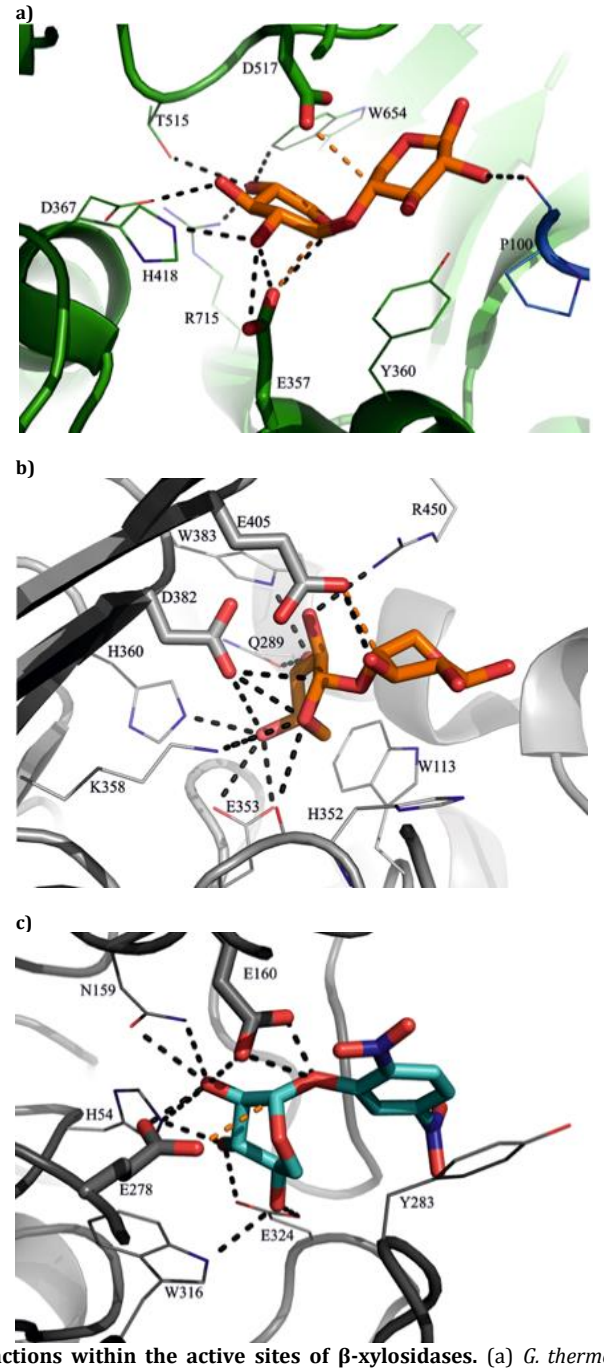


Figure 4.19. Interactions within the active sites of β-xylosidases. (a) *G. thermoglucosidasius* GH52. (b) GH120 (3VSU). (c) GH39 (2BFG). The substrate is shown in thick bonds (xylobiose with orange carbons and *p*NP-XP with cyan) and the interacting residues as sticks; hydrogen bonds are shown as black dashed lines and carbon-oxygen interactions less than 3.2 Å are shown in orange.

The solvent accessibility for the three structures was calculated using AREAIMOL, part of CCP4, and the results are shown in Table 4.9, while Figure 4.20 shows a surface representation of the enzyme's active site for clearer visualisation of the cleft.

Table 4.9. Solvent-accessibility analysis of the substrate bound in all three structures. Solvent accessible surface area calculated using AREAIMOL.

	GH52 (4C1P)	GH120 (3VSU)*	GH39 (2BFG) *
	O4B BXP = 0.0 Å²	O4B BXP = 0.0 Å²	C1 XYP = 0.0 Å²
	C4B BXP = 0.0 Å²	C4B BXP = 0.0 Å²	C2 XYP = 0.0 Å²
	C3B BXP = 0.0 Å²	C3B BXP = 0.0 Å²	C3 XYP = 0.0 Å²
	O3B BXP = 0.0 Å²	O3B BXP = 1.0 Å²	C4 XYP = 0.5 Å²
-1	C2B BXP = 0.0 Å²	C2B BXP = 0.6 Å²	C5 XYP = 0.0 Å²
	O2B BXP = 0.0 Å²	O2B BXP = 0.0 Å²	O2 XYP = 0.0 Å²
	C5B BXP = 0.0 Å²	C5B BXP = 0.0 Å²	O3 XYP = 0.0 Å²
	O5B BXP = 0.0 Å²	O5B BXP = 0.0 Å²	O4 XYP = 0.0 Å²
	C1B BXP = 0.0 Å²	C1B BXP = 0.0 Å²	O5 XYP = 0.0 Å²
	O4 A BXP = 0.0 Å²	O4 A BXP = 0.0 Å²	O1 ANX = 0.0 Å²
	C4 A BXP =0.0 Å²	C4 A BXP =0.0 Å²	C1 ANX = 0.0 Å²
	C3 A BXP = 0.0 Å²	C3 A BXP = 0.0 Å²	C2 ANX = 0.0 Å²
+1	O3 A BXP = 0.2 Å²	O3 A BXP = 0.0 Å²	N2 ANX = 0.0 Å²
	C5 A BXP = 0.0 Å²	O4B BXP = 4.6 Å²	O21 ANX = 0.0 Å²
	O5 A BXP = 0.0 Å²	C4B BXP = 11.1 Å²	O22 ANX = 0.0 Å²
	C1 A BXP = 6.4 Å²	C1 A BXP = 1.2 Å²	C3 ANX = 6.1 Å²
	C2 A BXP = 0.0 Å²	C2 A BXP = 4.5 Å²	C4 ANX = 13.1 Å²
	O2 A BXP = 7.1 Å²	O2 A BXP = 12.3 Å²	N1 ANX = 1.8 Å²
	O1 A BXP = 17.0 Å²	O1 A BXP = 44.7 Å²	O41 ANX = 27.7 Å²
			O42 ANX = 9.4 Å²
			C5 ANX = 0.3 Å²
			C6 ANX = 0.1 Å²

*Values for each atom correspond to the average of the entries from all the protein chains, 2 for 3VSU and 4 for 2BFG.

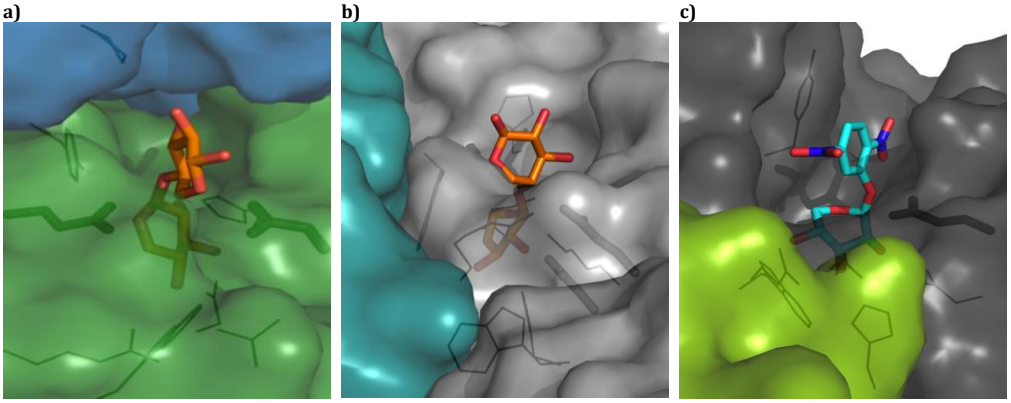


Figure 4.20. Active sites from different β-xylosidase families. A surface representation with 20% transparency, with the different monomers shown in different colours. The residues forming the cleft are shown in stick form and the catalytic residues are shown in thicker lines. (a) GH52. (b) GH120. (c) GH39.

The solvent-accessibility analysis of the substrate bound in all three structures reveals that the -1 xylose is completely buried, while the +1 xylose has variable levels of exposure to solvent (ANX in GH39 is the most accessible while GH52 is the most restricted). This lack of exposure of the -1 xylose to solvent would render the enzyme incapable of cleaving bulk xylan, thus explaining the lack of xylanase activity of these three β-xylosidases.

4.3.11. Comparison with similar structures.

As previously discussed, solving the *G. thermoglucosidasius* TM242 β-xylosidase-1 structure by molecular replacement was not possible as at the time of the structural studies presented in this chapter no structures available in the PDB database shared more than 20% sequence identity with the protein. However, once the structure had been solved, a search of the PDB and structural comparison using either the Protein Structure Comparison Service PDBeFold (Krissinel & Henrick, 2004) or the DALI server (Holm & Rosenström, 2010) showed that the TM242 β-xylosidase-1 shares striking structural similarities with other proteins, despite sharing less than 13% sequence identity. As all the PDBeFold results were also identified by DALI, only the DALI results are given in Table 4.10.

Table 4.10. Structures identified as similar to *G. thermoglucosidasius* TM242 β-xylosidase-1 by DaliLite V.3 (Holm & Rosenström, 2010) including only those structures that shared similarity across both domains.

PDB Code	Description	Organism	GH	% Id	Z score	RMSD Å	L ali	Nres
1UG9	Glucodextranase	*Arthrobacter globiformis*	15	8	25.9	4.7	379	1019
1LF6	Glucoamylase	*Thermoanaerobacterium thermosaccharolyticum*	15	8	24.5	5.4	491	674
3W7S	Uncharacterized protein Ygjk	*Escherichia coli K-12*	63	13	24.4	5.1	525	760
1V7V	Chitobiose phosphorylase	*Vibrio proteolyticus*	94	9	21.0	4.3	512	779
3CIH	Putative α-Rhamnosidase	*Bacteroides thetaiotaomicron*	78	10	20.9	3.6	338	714
4J5T	Mannosyl-oligosaccharide glucosidase	*Saccharomyces cerevisiae*	63	8	20.5	8.2	492	788
2CQS	Cellobiose phosphorylase	*Cellvibrio gilvus*	94	10	19.1	4.2	533	822
3RRS	Cellobiose phosphorylase	*Cellulomonas uda*	94	9	19.6	4.3	530	822
3QDE	Cellobiose phosphorylase	*Clostridium thermocellum*	94	11	19.4	4.3	524	811
1H54	Maltose phosphorylase	*Lactobacillus brevis*	65	8	17.3	5.5	504	754
3W5M	Putative rhamnosidase	*Streptomyces avermitilis*	-	9	17.5	3.4	326	1030
2RDY	Bh0842 Protein	*Bacillus halodurans*	-	7	12.9	4.7	494	787
2WW2	α-1,2-Mannosidase	*Bacteroides thetaiotaomicron*	92	6	11.8	5.7	385	737
2XSG	Ccman5	*Cellulosimicrobium cellulans*	92	7	10.3	4.9	376	765
2EAB	Alpha-Fucosidase	*Bifidobacterium bifidum*	95	10	11.6	4.8	495	888

Z = Z-score (statistical significance) of the best domain-domain alignment (distance, in standard deviations, between the observed alignment RMSD and the mean RMSD for random pairs of the same length, with the same or fewer gaps)
RMSD = root-mean-square deviation of Cα atoms in rigid-body superimposition
Lali = total number of structurally-equivalent residues
%Id = percentage of sequence identity over equivalenced positions

The structures identified as similar are mostly glycosidases, from the GH families 15, 63, 65, 78, 92, 94 and 95. Of these, just GH15 and GH65 have a clan assigned, both belonging to the clan GH-L that denotes an $(\alpha/\alpha)_6$ fold; this clan assignment would match the C-terminal domain of the *G. thermoglucosidasius* TM242 β-xylosidase-1. This identification and the very low sequence identity shared between these proteins with a very similar structure suggests that an ancestral protein contained this combination of structural domains, and has subsequently evolved to hydrolyse a wide range of glycosides to the extent that sequence identity is no longer discernible.

The structure of glucoamylase from *Thermoanaerobacterium thermosaccharolyticum* (PDB entry: 1LF9 (Aleshin *et al.*, 2003)) is the most similar protein of comparable amino acid sequence length to the TM242 β-xylosidase-1 while sharing just 8% sequence identity. The structural similarities of both protein folds are shown in Figure 4.21 and it can be seen that not only they are visually very similar but also that the substrates of both enzymes overlap, revealing that these proteins also share the position of their active sites. Furthermore, the classic position for an active site in an $(\alpha/\alpha)_6$ fold (in the groove at the centre of the barrel above the N-terminal ends of the inner helices) is conserved in all those structures containing both domains similar to the β-xylosidase-1. However, the dimeric nature of the *G. thermoglucosidasius* β-xylosidase is just conserved in half of the glycoside hydrolase families presented in Table 4.10, seven of them being monomeric and one trimeric.

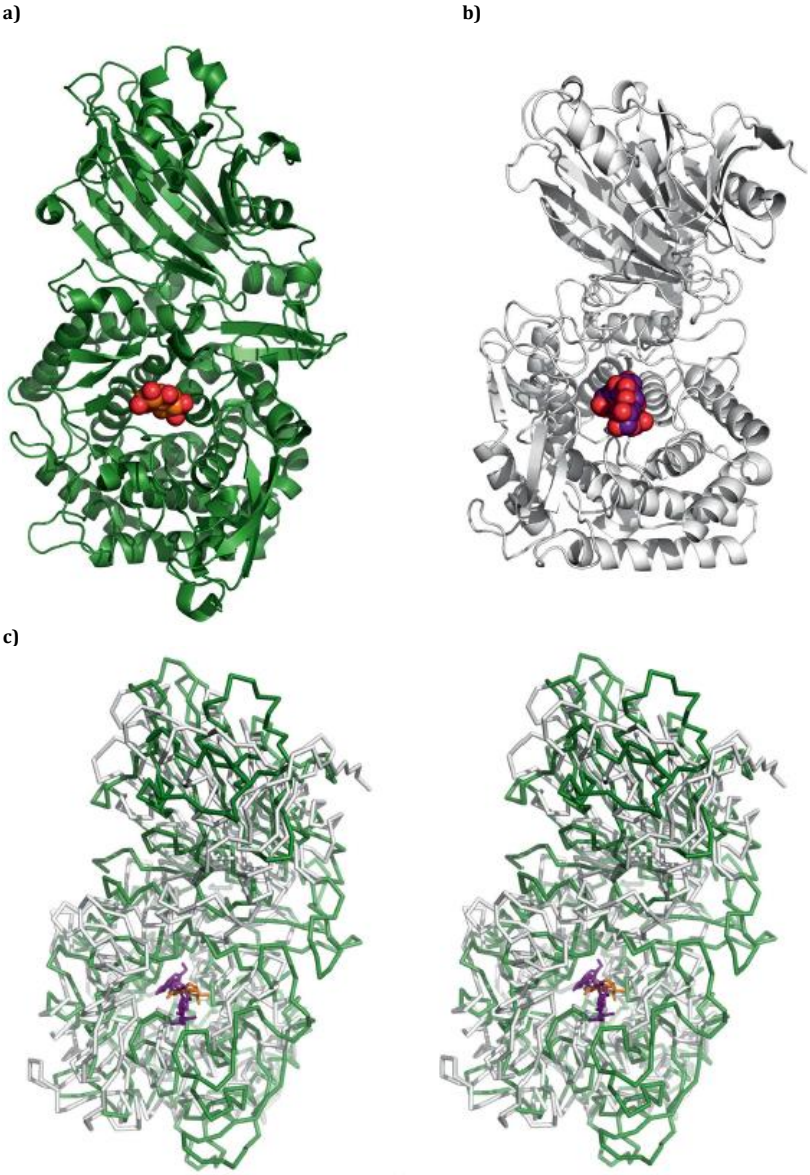


Figure 4.21. Structure alignment of *G. thermoglucosidasius* β-xylosidase (green) with a representative GH15 structure (white) chosen to be of similar size. (a) A cartoon representation of *G. thermoglucosidasius* TM242 β-xylosidase-1 with xylobiose in space-filling spheres. (b) A cartoon representation of *Thermoanaerobacterium thermosaccharolyticum* glucoamylase (PDB entry 1LF9 (Aleshin *et al.*, 2003)) with acarbose (inhibitor) bound. (c) A stereoview of the Cα traces of *G. thermoglucosidasius* β-xylosidase and the aligned GH15 structure, with the xylobiose in the β-xylosidase-1 shown as orange sticks and the acarbose inhibitor in the glucoamylase shown in purple.

4.4. CONCLUSIONS

A high-resolution (1.7 Å) crystal structure of β-xylosidase, along with a lower resolution structure (2.6 Å) of the enzyme-substrate complex, has been determined resulting in the first reported structure of a GH52 family member.

The novel β-xylosidase structure reveals that GH52 enzymes are composed of two domains, an N-terminal β-sandwich and a C-terminal $(\alpha/\alpha)_6$ barrel. The active site is located at the centre of the $(\alpha/\alpha)_6$ domain and the positioning of the xylobiose substrate is consistent with the family's proposed retaining mechanism.

Surprisingly, the fold of this enzyme shares structural similarities with other glycoside hydrolases, despite sharing less than 13% sequence identity, whilst it is completely different from the xylosidase structures in other CAZy families.

Both structures have been deposited in the PDB and given the access codes 4C1O and 4C1P, respectively, and this work has been published in *Acta Crystallographica Section D* (Espina *et al.*(2014); Appendix III).

CHAPTER 5: Identification and heterologous expression of genes encoding xylanases from closely related *Geobacillus* strains.

5.1. INTRODUCTION

As described in Chapter 1, hemicellulose, the second most abundant polysaccharide in Nature, is a heterogeneous group of branched and linear polymers that can comprise up to 35% of the lignocellulosic biomass (Saha, 2003). There are different types of hemicelluloses and they show differences in their structure and composition (*e.g.* backbone monomers, side chains and degree of substitutions) between different plant species and even in different cell types within one plant (Scheller & Ulvskov, 2010). However, in general, hemicelluloses are built up with pentoses (*e.g.* D-xylose, D-arabinose), hexoses (*e.g.* D-mannose, D-glucose, D-galactose, L-galactose, L-rhamnose, L-fucose) and/or uronic acids (*e.g.* D-glucuronic, D-4-O-methylgalacturonic and D-galacturonic acids) (Petrova & Ivanova, 2014).

Xylan, the most common hemicellulose, is a polysaccharide made of a homopolymeric backbone of β-1,4-glycosidically-linked D-xylopyranose units. Linear homoxylans have been isolated from guar seed husk, esparto grass, and tobacco stalks; however, most xylans are generally substituted with residues of arabinofuranose, glucuronic acid (or its 4-O-methyl ether) and acetic, ferulic, and *p*-coumaric acids, whose exact ratios vary accordingly to the xylan source (*e.g*. Birch wood xylan contains 89.3% xylose, 1% arabinose, 1.4% glucose, and 8.3% anhydrouronic acid, while rice bran neutral xylan contains 46% xylose, 44.9% arabinose, 6.1% galactose, 1.9% glucose, and 1.1% anhydrouronic acid, and wheat contains 65.8% xylose, 33.5% arabinose, 0.1% mannose, 0.1% galactose, and 0.3% glucose) (Saha, 2003; Scheller & Ulvskov, 2010).

As outlined in Chapter 3 (and detailed in Chapter 1), total biodegradation of hemicellulose to its simple sugars involves the synergistic action of several enzymes known as hemicellulases. Among these, xylanases (EC 3.2.1.8) and β-xylosidases (EC 3.2.1.37) are crucial for the depolymerisation of the xylan backbone, as xylanases randomly attack the main chain of xylans and cleave the internal linkages to yield xylo-oligosaccharides, which are further hydrolysed to xylose by the action of β-xylosidases.

It has been found that many xylanases do not cleave glycosidic bonds between xylose units that are substituted, so before the xylan backbone can be completely hydrolysed, the side chains must be cleaved by auxiliary hemicellulases (Saha *et al.*, 2009). However, in some cases these debranching enzymes will only remove side chains from xylo-oligosaccharides, requiring the action of xylanases to partially hydrolyse the xylan backbone before the side chains can be cleaved (Saha *et al.*, 2009); therefore, a pool of enzymes containing different hemicellulases is required in order to completely and effectively breakdown heteroxylans.

The investigation of glycoside hydrolases from *G. thermoglucosidasius* TM242 has been presented throughout the previous chapters of this thesis, where genes encoding three hemicellulases (two β-xylosidases and one α-arabinofuranosidase) have been cloned and expressed, and the recombinant enzymes characterised. Chapter 4 presented the first crystal structure of the β-xylosidase from the family GH52. However, the search of the TM242 genome sequence revealed this microorganism lacks genes encoding xylanases that are required for thorough hemicellulose degradation.

On the other hand, closely related *Geobacillus* strains possess at least one xylanase-encoding gene in their genome (Anand *et al.*, 2013; Canakci *et al.*, 2012; Khasin *et al.*, 1993; Liu *et al.*, 2012); thus, it can be suggested that the missing xylanases in *G. thermoglucosidasius* TM242 could be an evolutionary indication of the bacterium living in symbiotic communities and relying on other microorganisms to break down large carbohydrate polymers. The wild-type microorganism, *G. thermoglucosidasius* NCIMB 11955, was isolated from a compost heap environment that would contain a wide variety of other thermophilic microorganisms forming a consortium where strains deficient in genes or pathways required for hemicellulose degradation and utilisation may be complemented by other members of the consortium. Therefore, it can be hypothesised that 11955, in its ability to utilise a wide range of sugars, just used the oligosaccharides that other microoorganisms produced, with no need to express its own xylanases.

The efficient expression of heterologous genes encoding glycoside hydrolases (with and without predicted signal peptides sequences) in other *Geobacillus* strains (*e.g.* α-amylase from *G. stearothermophilus* expressed in *G. kaustophilus* HTA426 (Suzuki *et al.*, 2013) and Cel5A endoglucanase from *Thermotoga maritima* expressed in *G. thermoglucosidasius* 11955 (Bartosiak-Jentys *et al.*, 2013) has been achieved. This

suggests it might be possible to incorporate the missing xylanase-encoding genes into TM242, in order to enhance its degradative ability and the biomass utilisation by this microorganism.

This chapter describes the cloning and heterologous expression (in *E. coli* as well as in *G. thermoglucosidasius* TM242) of four different xylanase-encoding genes from closely related thermophilic microorganisms (two from *G. thermoglucosidasius* C56-YS93, one from *G. kaustophilus* and one from *G. stearothermophilus/G. thermoleovorans*). The assessment of the xylanolytic activity of the expressed enzymes is also reported.

5.1.1. Cloning, recombinant expression and characterisation

The cloning of the closely related xylanase-encoding genes in pET-28a(+) was achieved as previously described in Chapter 3 (Section 3.1.1 and 3.2.5). For the cloning and heterologous expression in *G. thermoglucosidasius* TM242, the expression vector of choice was pUCG4.8 (Appendix I.c), which is a modified version of the *E. coli-Geobacillus* sp. shuttle vector pCEX3.1.2 (Bartosiak-Jentys *et al.*, 2013). Some of the unique features of the expression vector include the *E. coli* pMB1 origin of replication and MCS of pUC18, the *Geobacillus* sp. origin of replication from the *G. stearothermophilus* cryptic plasmid pBST1, a thermostable variant (TK101) of the kanamycin nucleotidyltransferase gene (*knt*) and a synthetic, constitutive PUp2n38promoter. The promoter was obtained by Dr Ben Reeve, Imperial College London, and its sequence was modified from the region upstream of the Uracil phosphoribosyltransferase (an enzyme involved in the pathway employed to salvage uracil) in *G. thermoglucosidasius* NCIMB 11955 that was shown to be constitutively active and to offer moderate expression, with consensus -10 and -35 boxes. Different from pCEX3.1.2, the expression vector pUCG4.8 has the advantage of having the addition of an ampicillin resistance gene (*bla*) as a second antibiotic marker for easier selection of *E. coli* clones (obtained by Dr Steven Bowden, University of Bath).

The recombinant proteins were expressed in *E. coli* from their genes cloned in pET-28a(+), and were then purified as previously described in Chapter 3 (Section 3.1.2). Cell extracts of *G. thermoglucosidasius* TM242 harbouring pUCG4.8-xylanase vectors and their concentrated supernatants were analysed without purification. The heterologously-expressed xylanases (in both *E. coli* and TM242) were then assayed by means of two discontinuous xylanolytic activity assays: (a) plate assays of minimal media containing 1% of different xylan sources, where xylan hydrolysis was revealed

by the appeareance of a halo, which could be more visible by staining with Congo Red dye that interacts with polysaccharides containing 1,4, 1,3 β-D-glucosidic linkages (Teather & Wood, 1982); and (b) 3,5-dinitrosalicylic acid (DNS) assay, where the enzyme activity was assayed by measuring the amount of xylose equivalent liberated from xylan hydrolysis using DNS (an aromatic compound that reacts with reducing sugars to form 3 amino-5-nitrosalicylic acid) (Miller, 1959).

5.2. MATERIALS AND METHODS

5.2.1. Identification of xylanase genes in closely related Geobacillus strains

Three *Geobacillus* strains, closely related to *G. thermoglucosidasius* TM242 were available for study: *G. thermoglucosidasius* C56-YS93, *Geobacillus* sp. CAMR12739 and *Geobacillus* sp. CAMR5420.

The identification of xylanase-encoding genes in *G. thermoglucosidasius* C56-YS93 was done by comparing its genome (which is publicly available) with known nucleotide sequences of carbohydrate-degrading enzymes mainly involved with cellulose and hemicellulose hydrolysis. The search was predominantly restricted to *Bacillus* related species and sequences were obtained from 28 different GH families (1, 3, 5, 6, 7, 8, 9, 10, 11, 12, 16, 18, 19, 26, 30, 39, 43, 44, 45, 48, 51, 52, 54, 61, 62, 74, 116, 120) in the CAZy database (Lombard *et al.*, 2014). The online comparison was carried out using the private database ERGO (Overbeek *et al.*, 2003), which also provided the gene annotation and predicted open reading frame for each sequence.

The two isolated CAMR12739 and CAMR5420 strains were previously shown to possess high xylanolytic activity in studies performed by Carolyn Williamson and Chris Vennard at the CER, and their genome sequences were determined by the Research Group of Professor Don Cowan, University of Pretoria, South Africa (De Maayer *et al.*, 2014). The nucleotide sequences of the xylanase- and β-xylosidase-encoding genes found in both *Geobacillus* strains were kindly shared by Prof Don Cowan (personal communication).

5.2.2. Bioinformatic analysis of TM242's glycoside hydrolases

The *Geobacillus* xylanases sequences were analysed as described in 3.2.2.

5.2.3. Primer design

Forward and reverse primers were designed to amplify and subsequently clone the xylanase-encoding genes into two different expression vectors: pET-28a(+) and pUCG4.8. Forward and reverse primers were designed to be identical to the desired 5' region of the gene of interest and its 3' end, respectively. The primers also incorporated an appropriate restriction site (not present in the gene of interest) at the desired terminus of the DNA sequence. The length of each sequence (24-32 bases) was determined based on similar melting temperatures (T_m). The primers were all synthesised by Invitrogen Ltd (Paisley, UK) or Eurofins MWG Operon (Ebersberg, Germany).

5.2.3.1. For E. coli cloning

The nucleotide sequences of four xylanase-encoding genes (two from *G. thermoglucosidasius* C56-YS93, one from CAMR12739 and one from CAMR5420) were used to design the specific primers shown in Table 5.1. These included the restriction sites for NheI (forward) or XhoI (reverse) at each 5' end, as required to clone the amplified genes into the expression vector pET-28a(+) including an N-terminal His tag. Predicted signal peptides sequences were not incorporated in order to ensure solubility, and therefore start codons were introduced into each forward primer.

Table 5.1. Primers used to amplify xylanases-encoding genes from *Geobacillus* strains and clone them into the vector pET-28a(+) for expression in *E. coli*. The restriction sites introduced into each primer are underlined. Extra bases introduced to favour the cleavage are shown in lower case. The start codons introduced are shown in bold.

Specific Gene	Name	Sequence	Length	Tm
C56-YS93 extracellular xylanase	XC56-F	5' g<u>GCTAGC</u>**ATG**GCAGATACGGCTTCCTAT3'	28 b	70°C
	XC56-R	5' cc<u>gCTCGAG</u>TTATTTATGATCGATAATGGC 3'	30 b	69°C
C56-YS93 intracellular xylanase	IXC56-F	5' g<u>GCTAGC</u>**ATG**AGCAGCTCGCTTCCTT 3'	26 b	71°C
	IXC56-R	5' c<u>CTCGAG</u>TTACCCGTTGACGACCCTC 3'	26 b	61°C
CAMR12739 extracellular xylanase	X12739-F	5' cg<u>GCTAGC</u>**ATG**AAGAATGCAGATTCCTAT 3'	29 b	68°C
	X12739-R	5' cc<u>gCTCGAG</u>TTAGAAAGAGGATGCC 3'	25 b	68°C
CAMR 5420 extracellular xylanase	X5420-F	5' cta<u>GCTAGC</u>**ATG**AAAACGGAACCATCGTACG 3'	31 b	70°C
	X5420-R	5' c<u>CTCGAG</u>CTATTTGTGGTCGATAATAGCCC 3'	30 b	70°C

5.2.3.2. For Geobacillus cloning

The nucleotide sequences of the three genes encoding extracellular xylanases were used to design the specific primers shown in Table 5.2. These included restriction sites for BstBI (forward) or SacI (reverse) at each 5' end, as required to clone the amplified genes (including their predicted signal peptide sequences) into the expression vector pUCG4.8.

Table 5.2.Primers used to amplify xylanases-encoding genes from *Geobacillus* strains and clone them into the vector pUCG4.8 for expression in *G. thermoglucosidasius*TM242.The restriction sites introduced into each primer are underlined.

Specific Gene	Name	Sequence	Length	Tm
C56-YS93 extracellular xylanase	TXC56-F	5' <u>TTCGAA</u>ATGCGGAACGTTTTACGC 3'	24 b	68°C
	TXC56-R	5'<u>GAGCTC</u>TTATTTATGATCGATAATGGCCCAATAG 3'	34 b	68°C
CAMR12739 extracellular xylanase	TX12739-F	5'<u>TTCGAA</u>ATGAGGGGGAATAGTATGCG 3'	26 b	67°C
	TX12739-R	5' <u>GAGCTC</u>TTAGAAAGAGGATGCCTCGTTG 3'	28 b	67°C
CAMR 5420 extracellular xylanase	TX5420-F	5' <u>TTCGAA</u>ATGATCGTTGGATTCTCGTTTATG 3'	30 b	69°C
	TX5420-R	5' <u>GAGCTC</u>CTATTTGTGGTCGATAATAGCCC 3'	29 b	67°C

5.2.4. Amplification of Geobacillus xylanases

The three *Geobacillus* strains were grown aerobically at 60°C for 16 h in 2SPYNG media and their genomic DNAs were isolated using standard techniques described in Sambrook & Russell (2001). These genomic DNAs were used as templates for PCR reactions previously described in 2.2.1.1.

5.2.5. Cloning of xylanases

5.2.5.1. Into E. coli expression vector pET-28a(+)

Following PCR amplification, the xylanase-encoding genes were purified purified, A-tailed and ligated into pGEM®-T Easy vector as described in 2.2.3, 2.2.4.1 and 2.2.4.2 respectively. The ligations were ethanol-precipitated as described in 2.2.4.3 and transformed by electroporation into *E. coli* JM109 electrocompetent cells following 2.2.4.4. The positive clones (harbouring pGEM®-T Easy vector containing xylanase-encoding genes) were selected as described in 2.2.4.5. The correct plasmids were purified following 2.2.5, digested with NheI and XhoI restriction enzymes and then ligated into the NheI/XhoI sites of the expression vector pET-28a(+) multiple cloning site, following the methods previously described in 2.2.7.1 and 2.2.7.2 respectively. These ligations were then ethanol-precipitated and transformed by electroporation into *E. coli* JM109 electrocompetent cells as before and the positive clones (harbouring pET-28a(+) containing xylanase-encoding genes) were selected as described in 2.2.7.3. Finally, the correct plasmids were purified as before, sequenced following 2.2.6 and then transformed by heat shock into chemically competent *E. coli* BL21(DE3) cells as described in 2.2.7.4.1. To preserve bacteria harbouring expression vectors with verified insert sequences, glycerol stocks were prepared as described in 2.1.2.3 and stored at -80°C.

5.2.5.2. Into Geobacillus expression vector pUCG4.8

Following PCR amplification, the xylanase-encoding genes were purified purified, A-tailed and ligated into pGEM®-T Easy vector as described in 2.2.3, 2.2.4.1 and 2.2.4.2 respectively. The ligations were ethanol-precipitated as described in 2.2.4.3 and transformed by electroporation into *E. coli* JM109 electrocompetent cells following 2.2.4.4. The positive clones (harbouring pGEM®-T Easy vector containing xylanase-encoding genes) were selected as described in 2.2.4.5. The correct plasmids were purified following 2.2.5, digested with BstBI and SacI restriction enzymes and then ligated into the ClaI/SacI sites of the expression vector pUCG4.8 multiple cloning site, following the methods previously described in 2.2.7.1 and 2.2.7.2 respectively. These ligations were then ethanol precipitated and transformed by electroporation into *E. coli* JM109 as before and the positive clones (harbouring pUCG4.8 containing xylanase-encoding genes) were selected as described in 2.2.7.3. Finally, the correct plasmids were purified as before, sequenced following 2.2.26 and then transformed by electroporation into *G. thermoglucosidasius* TM242 electrocompetent cells as described

in 2.2.7.4.2. To preserve bacteria harbouring expression vectors with verified insert sequences, glycerol stocks were prepared as described in 2.1.2.3 and stored at -80°C.

5.2.6. Protein expression

The xylanases were heterologously expressed in *E. coli* and *G. thermoglucosidasius* TM242 following the methods described in 2.3.1.2 and 2.3.2, respectively.

5.2.7. Protein purification

The recombinant His-tagged xylanases expressed in *E. coli* were purified by small-scale nickel affinity chromatography as described in 2.3.5.1.1, while the cell extract of *G. thermoglucosidasius* TM242 and its concentrated supernatant were analysed directly without enzyme purification.

5.2.8. Discontinuous xylanolytic activity assays

5.2.8.1. Plate Assays

Solid plates of USM media containing 2% agar and 1% xylan (Sigma-Aldrich) (from birchwood, beechwood or oat spelt according to the manufacturer availability) were prepared to detect xylanolytic activity. After the agar set, a sterilised cork borer was used to make wells if needed. Depending on the experiment, colonies or 40 µL of cell extract (purified protein of similar concentration or concentrated supernatant) were incubated overnight at 60°C and xylan degradation was indicated by the appearance of a halo.

In some cases, in order to make the clearance zone more visible, the plates were stained with 0.1% (w/v) Congo Red (Sigma-Aldrich) aqueous solution for 15 min. Finally, the plates were washed three times (20 min each) with 1 M NaCl to destain the unbound regions, revealing xylan hydrolysis.

5.2.8.2. 3,5-Dinitrosalicylic acid(DNS) Assays

Xylanase activity was also assayed by measuring the amount of reducing sugars as xylose equivalents liberated by xylan hydrolysis using DNS as described by (Miller, 1959). The substrate used was 1% (w/v) birchwood xylan (Sigma-Aldrich) in McIlvaine buffer pH 7.0, and was prepared as described by Bailey *et al.* (1992). The reaction mixture, containing 450 µL of the substrate and 50 µL of appropriately diluted enzyme/cell extract volume in McIlvaine buffer pH 7.0, was incubated at 60°C in a heating block for different lengths of time (1-15 min). The reaction was stopped by the

addition of 750 µl DNS reagent (1% (w/v) DNS, 1% (w/v) NaOH and 0.05% (w/v) Na_2SO_3) before heating the mixture for 10 min. Then 250 µl of 40% (w/v) Rochelle salts ($KNaC_4H_4O_6 \cdot 4H_2O$) were added to the samples and were allowed to cool for 5 min on ice. A control was run simultaneously that contained all the reagents except the enzyme to determine the presence of reducing sugars in the substrate. The reactions were centrifuged for 2 min at 14,000g and 150 µL were transferred to a 96 well microplate (Corning). Finally, the absorbance was measured at 540 nm in a Biotek Synergy HT plate reader.

A 10 mM xylose stock solution was used to make standards of known concentrations (0-10mM). This standard curve was used to calculate the amount of reducing sugar as xylose equivalents liberated from the samples. One unit (U) of xylanase activity was defined as the amount of enzyme that catalyses the release of 1 µmol of reducing sugar as xylose equivalents per minute, based on the xylose calibration curve, under the specified assay conditions.

5.3. RESULTS AND DISCUSSION

5.3.1. Identification of xylanase genes in Geobacillus spp.

The bioinformatic analysis of the *G. thermoglucosidasius* C56-YS93 genome resulted in the identification of thirteen genes encoding enzymes presumed to be involved with cellulose or hemicellulose hydrolysis (with seven sharing more than 85% sequence similarity with the ones identified in TM242), of which two genes were annotated to encode xylanases (and one β-xylosidase from GH family 39) not present in *G. thermoglucosidasius* TM242. The genome sequences of the two isolated strains, CAMR12739 and CAMR5420, also revealed two xylanase encoding genes in each (and one β-xylosidase - from GH43 in the case of CAMR5420 and GH39 for CAMR 12739).

The amino acid sequences of the six different xylanases were then analysed with SignalP4.0 and the CDD to check for signal peptides and the presence of carbohydrate binding modules. The results indicated that for each strain, one of the two xylanases possesses a predicted signal peptide sequence, while the other is annotated as an intracellular enzyme. The six different xylanases were found to belong to GH family 10 with no CBMs.

In addition, a PDB search with the amino acid sequence of the three mature extracellular xylanases revealed the GH10 extracellular xylanases, XT6, from *Geobacillus stearothermophilus* T-6 (PDB entry: 1HIZ; structure obtained at 2.4 Å resolution) to be the most similar sequence in the database (Alignment: E-value: 0.0; Score: 732.635 bits (1890); Identities: 93-99%; Positives: 96-99%; Gaps: 0%); while the PDB search with the amino acid sequence of the three intracellular xylanases revealed the GH10 intracellular xylanases, IXT6, also from *G. stearothermophilus* T-6 (PDB entry: 2Q8X; structure obtained at 1.45 Å resolution) to be the most similar sequence in the database (Alignment: E-value: 9.53408E-178-0.0; Score: 620.157-682.559 bits (1598-1760); Identities: 88%-98%; Positives: 94%-99%; Gaps: 0%).

5.3.2. Amplification, cloning and recombinant expression in E. coli

The three genes encoding extracellular xylanases (one from each strain), plus the intracellular one from *G. thermoglucosidasius* C56-YS93, were selected for further study: cloning and recombinant expression in *E. coli*.

The four gene sequences were correctly amplified using their respective *Geobacillus* genomic DNA as template. Specific PCR products of the approximate expected size were obtained as shown in Figure 5.1. The bands were excised from the gel, A-tailed and cloned into pGEM®-T Easy, and then subcloned into the expression vector pET-28a(+).

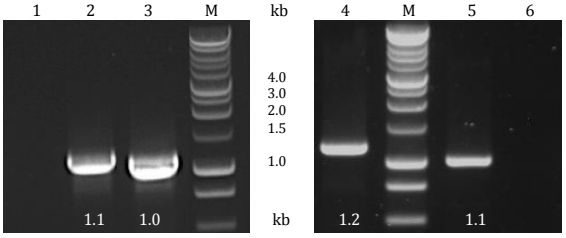


Figure 5.1. PCR amplification of xylanase-encoding genes from different *Geobacillus* genomic DNA. Amplicons are shown on a 0.8% agarose gel corresponding to: (2) XC56 extracellular xylanase from *G. thermoglucosidasius* C56-YS93 (without its signal peptide), (3) IXC56 intracellular xylanase from *G. thermoglucosidasius* C56-YS93, (4) X12739 extracellular xylanase from CAMR12739 (without its signal peptide), (5) X5420 extracellular xylanase from CAMR5420 (without its signal peptide), and (M) marker lanes containing a 1 kb ladder (Promega). Negative controls (no DNA) are shown in lanes 1 and 6. The expected sizes in kb for each of the PCR products are shown below each band.

Sequencing confirmed the cloned gene sequences were correct and the vectors harbouring the xylanase-encoding genes were then transformed into *E. coli* BL21(DE3) and grown in Overnight Express™ medium. Successful expression of each xylanase in soluble form was achieved and the enzymes were purified from the soluble cell extract by small-scale IMAC purification. The SDS-PAGE gels with each fraction are presented in Figure 5.2-5.5.

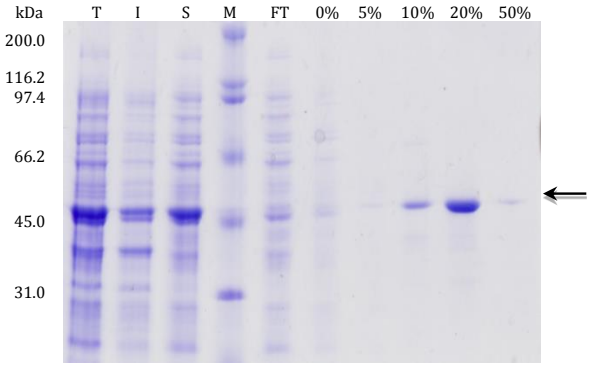


Figure 5.2. SDS-PAGE gel of the extracellular xylanase (XC56) from *G. thermoglucosidasius* C56-YS93 purified by IMAC. Lane T corresponds to the total cell extract, I is the insoluble fraction of the cell extract, S is the soluble fraction of the cell extract, and FT is the flow through after loading the soluble fraction onto the IMAC column. The following lanes are labelled by the percentage of His-elute buffer used to elute proteins from the column. The standard protein markers are shown in lane M and XC56 is highlighted with an arrow (predicted M_r = 47.2kDa).

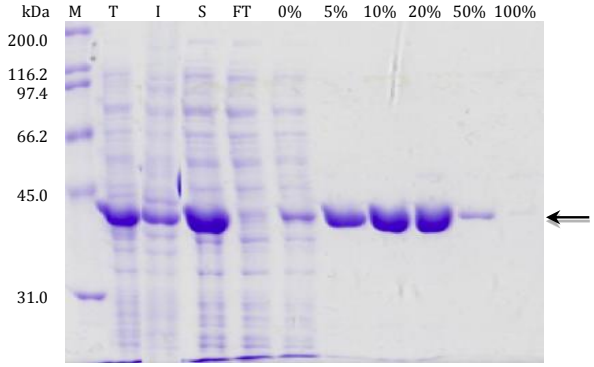


Figure 5.3. SDS-PAGE gel of the intracellular xylanase (IXC56) from *G. thermoglucosidasius* C56-YS93 purified by IMAC. Lane T corresponds to the total cell extract, I is the insoluble fraction of the cell extract, S is the soluble fraction of the cell extract, and FT is the flow through after loading the soluble fraction onto the IMAC column. The following lanes are labelled by the percentage of His-elute buffer used to elute proteins from the column. The standard protein markers are shown in lane M and IXC56 is highlighted with an arrow (predicted M_r = 41.1kDa).

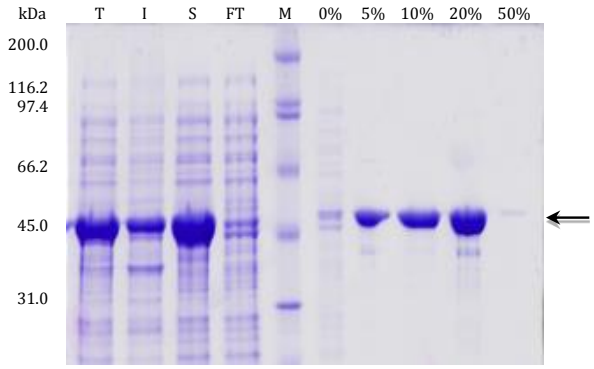


Figure 5.4. SDS-PAGE gel of the extracellular xylanase (X12739) from isolated CAMR12739 purified by IMAC. Lane T corresponds to the total cell extract, I is the insoluble fraction of the cell extract, S is the soluble fraction of the cell extract, and FT is the flow through after loading the soluble fraction onto the IMAC column. The following lanes are labelled by the percentage of His-elute buffer used to elute proteins from the column. The standard protein markers are shown in lane M and X12739 is highlighted with an arrow (predicted M_r = 49.2kDa).

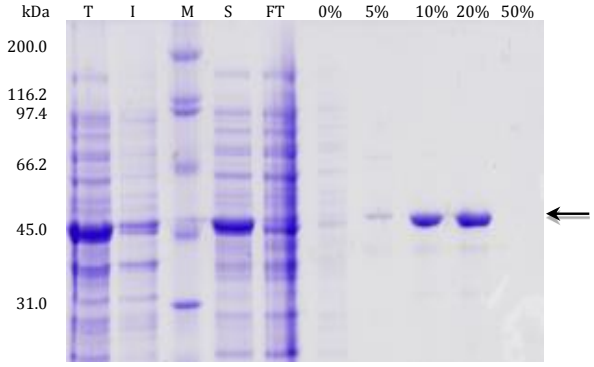


Figure 5.5. SDS-PAGE gel of the extracellular xylanase(X5420) from isolated CAMR5420purified by IMAC. Lane T corresponds to the total cell extract, I is the insoluble fraction of the cell extract, S is the soluble fraction of the cell extract, and FT is the flow through after loading the soluble fraction onto the IMAC column. The following lanes are labelled by the percentage of His-elute buffer used to elute proteins from the column. The standard protein markers are shown in lane M and X5420 is highlighted with an arrow (predicted M_r = 46.6kDa).

Through comparison with the standard proteins from the broad range marker (Biorad), the size of the recombinant xylanases were determined to be 46.0 kDa (XC56), 40.2 kDa (IXC56), 50.4 kDa (X12739) and 47.3 kDa (X5420), which agrees with their respective expected size of 47.2, 41.1, 49.2 and 46.6 kDa calculated by the ProtParam tool (Gasteiger *et al.*, 2003).

5.3.3. Structurally-based comparison of the different xylanases

As previously mentioned the three different extracellular xylanases studied in this chapter share the highest sequence similarity with the extracellular xylanase (XT6) from *G. stearothermophilus* T-6. The XT6 crystal structure as described by Teplitsky *et al.* (2004) is a monomer with a classical $(\beta/\alpha)_8$ barrel fold characteristic of GH10 xylanases. The shape of the barrel is elliptical in cross-section, with the major axis running between β-strands 1 and 5 of the $(\beta/\alpha)_8$ domain. As commonly observed, the top of the barrel is higher at the end points of the long elliptical cross-section axis and lower at the end points of the short elliptical axis, resulting in a bowl-shaped cavity also referred to as a 'salad-bowl'(Figure 5.6).

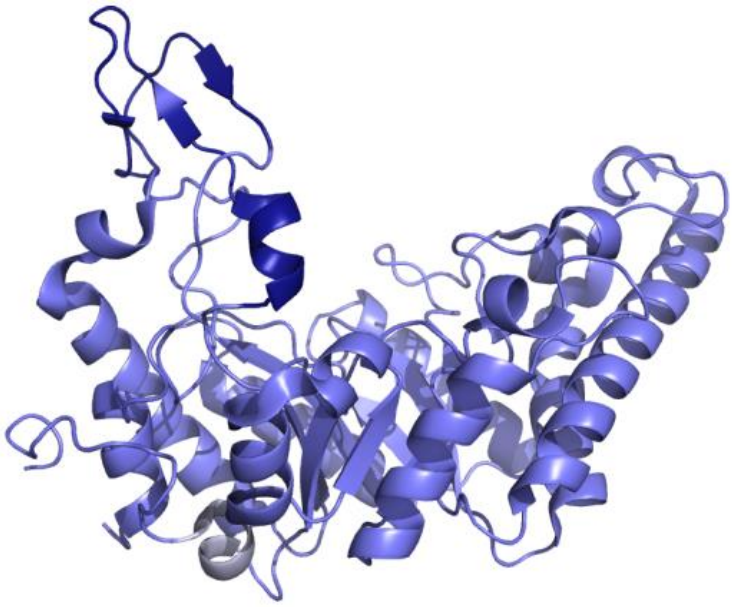


Figure 5.6. Cartoon representation of GH10 extracellular xylanase (XT6) from *G. stearothermophilus* (PDB entry: 1HIZ). The protein structure is monomeric with a $(\beta/\alpha)_8$ barrel fold with an additional N terminal α-helix (shown in lighter colour) and an atypical subdomain composed of a short α-helix followed by three short antiparallel β-strands connected by a long bended loop at the C-terminus (shown in darker colour) (Teplitsky *et al.*, 2004).

XT6 also possess an additional amino terminal α-helix and a unique subdomain (shown in lighter and darker colour, respectively, in Figure 5.6) composed of a short α-helix followed by three short antiparallel β-strands connected by a long bent loop and located at the C-terminal end ('top') of the α/β-barrel of XT6, between the classical eighth β-strand (β_{8a}) and the eighth α-helix (α_{8b}). The rim along the elliptical long axis of the top cross-section observed for XT6 is higher in comparison to other xylanases, which mostly arises from the existence of the additional subdomain that spreads out of the molecule away from the centre of the β-barrel and generally above the opening of the expected substrate-binding area (Teplitsky *et al.*, 2004). The exact role of this subdomain remains unknown, but from its location in the vicinity of the active site, the authors state it is likely to be involved in specific substrate binding and it is tempting to speculate that its flexibility could act as a `lid'over the active-site cleft.

The catalytic active site of XT6 is located on a long and open groove on the protein surface at the C-terminal end of the β-barrel; its structure and dimensions correlate well with those of a xylan chain, which is consistent with the 'endo' type of action on the large polysaccharide backbone and with the location and the shape of the corresponding substrate-binding grooves found in homologous xylanases of family GH10 (Teplitsky *et al.*, 2004). The active site of GH10 xylanases usually forms an open cleft surrounded by loops that are rich in aromatic residues, generally involved in substrate binding; in XT6 the active site groove is exposed to the solvent and is lined by an array of aromatic and hydrophilic residues. Figure 5.7 shows a surface representation (with 20% transparency) of the crystal structure of XT6 in complex with its substrate: cleaved xylopentaose (PDB entry: 1R87).

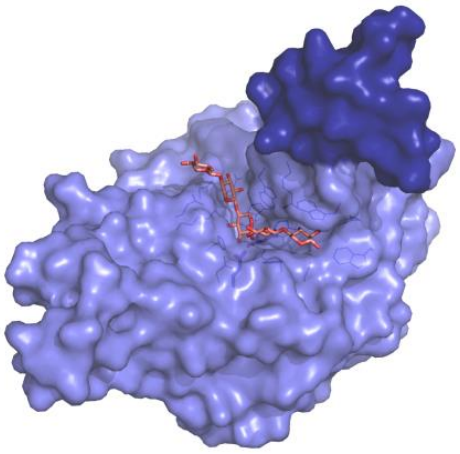


Figure 5.7. XT6 active-site groove in complex with its xylopentaose substrate shown in stick form (PDB entry: 1R87).A surface representation with 20% transparency; the enzyme's atypical subdomain is shown in darker colour.

Two carboxylic amino acids were identified as the catalytic residues: Glu159 (the acid/base), located in a β-bulge in the C-terminal part of the β4 strand, and the Glu265 (the nucleophile), located in a β-bulge on the β7 strand (Teplitsky *et al.*, 2004). This position of the catalytic carboxylates is characteristic of glycosidases belonging to the superfamily 4/7 (Jenkins *et al.*, 1995), and the distance between them (5.5 Å) is consistent with the GH10 retaining mechanism of hydrolysis (previously described in 4.3.9.1).

Zolotnitsky *et al.* (2004) published the results of an Isothermal Titration Calorimetry (ITC) analysis of the two different *G. stearothermophilus* T-6 xylanases. The study helped to elucidate the thermodynamics of xylosaccharide binding, and provided valuable information concerning the enzyme–substrate interactions and the number of sugar binding sites of these enzymes (Zolotnitsky *et al.*, 2004). The active site and the amino acids responsible for the catalysis, substrate binding and stacking interactions are shown in Figure 5.8.

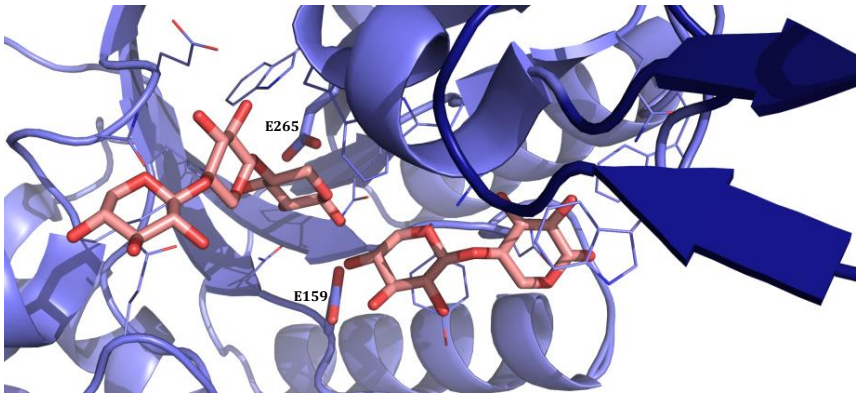


Figure 5.8. XT6 active-site cleft in complex with cleaved xylopentaose (PDB entry: 1R87). Residues forming the cleft are shown in stick form and the carboxylic catalytic residues E159 and E265 are shown in thicker lines. The enzyme's atypical subdomain is shown in darker colour.

As in other xylanases, the substrate-binding subsites of XT6 are mainly defined by aromatic residues taking part in specific binding interactions (via hydrophobic and stacking interactions) with corresponding substrate sugar units. According to Zolotnitsky *et al.* (2004), the values obtained by ITC combined with the structural data available, suggested there are six specific sugar binding subsites in this enzyme and these were designated as -3 to +3 (from the non-reducing end to the reducing end of the sugar substrate), with the cleavage occurring between the -1 and +1 subsites (Zolotnitsky *et al.*, 2004). Based on both the crystal structure and the thermodynamic

binding measurements of xylo-saccharides in solution, a few specific amino acid residues of XT6 were associated with each particular substrate-binding subsite: Gln102(-3), Trp316(-2), Trp324(-1), Tyr203(+1), Trp273(+2) and Trp241(+3) (Solomon *et al.*, 2007). A detailed scheme of the residues involved and XT6 binding subsites is shown in Figure 5.9.

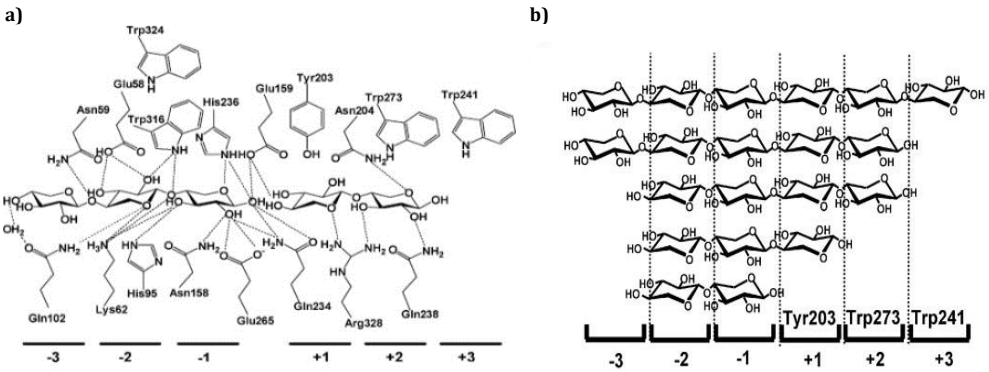


Figure 5.9. Substrate positional binding in XT6. (a) Schematic representation of the hydrogen bonding network and stacking interactions in the XT6 binding site. (b) The suggested positional binding of xylosaccharides to the binding subsites of XT6, the aromatic amino acids from +1, +2, +3 subsites form stacking interactions with the substrate. Images obtained from Zolotnitsky *et al.*, 2004.

On the other hand, the intracellular xylanases from the *Geobacillus* strains studied share the highest sequence similarity with the intracellular xylanase (IXT6) from *G. stearothermophilus* T-6. The IXT6 crystal structure, as described by Solomon *et al.* (2007), is also a monomer with the classical $(\beta/\alpha)_8$ barrel fold characteristic of GH10 xylanases (Figure 5.10).

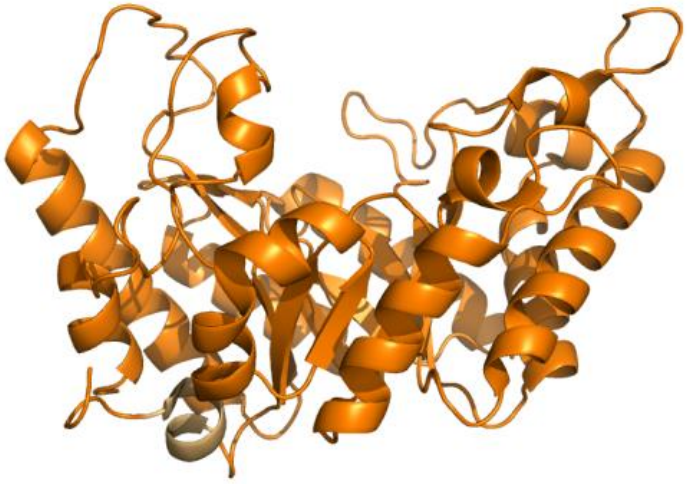


Figure 5.10. Cartoon representation of GH10 intracellular xylanase (IXT6) from *G. stearothermophilus* (PDB entry: 2Q8X). The protein structure is monomeric with a $(\beta/\alpha)_8$ barrel fold with an additional N-terminal α-helix (shown in lighter colour) (Solomon *et al.*, 2007).

Interestingly, both the *G. stearothermophilus* T-6 extracellular and intracellular xylanases belong to the same GH family and share 41% sequence identity (58% sequence similarity). Figure 5.11 shows the superposition of both T-6 xylanase structures, which with few exceptions are practically identical. Most of their differences are located at the 'top' of both molecules, around the upper rim of the 'salad-bowl', which is a relatively flexible area as might be expected from a region used for initial interactions with incoming substrates. The unique C-terminal subdomain of XT6 represents the most pronounced difference between the two structures as it is not present in IXT6.

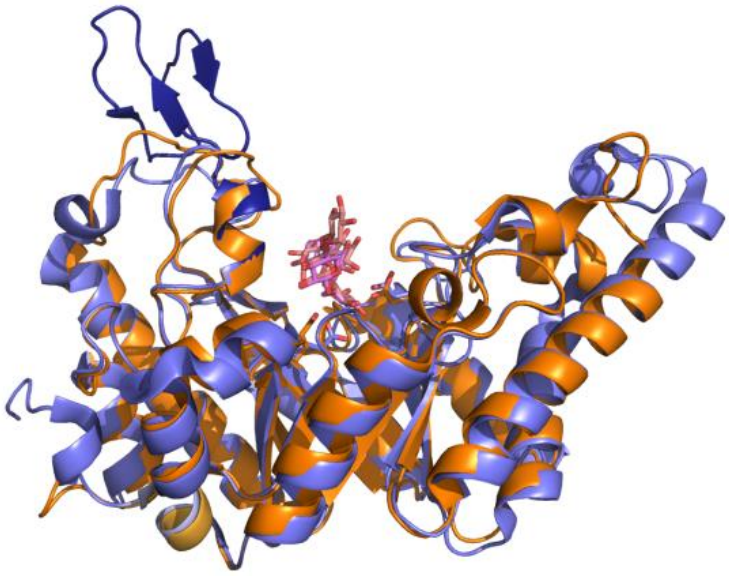


Figure 5.11. Structure alignment of *G. stearothermophilus* xylanases. A cartoon representation of the extracellular xylanase XT6 shown in purple (with its atypical subdomain shown in darker colour) and the intracellular xylanase IXT6 shown in orange. The xylo-oligomer substrates are shown as pink sticks.

While solving the IXT6 protein structure, Solomon *et al.* (2007) found thirteen glycerol molecules (originating from the cryoprotectant solution) bound in the crystal, with two of them tightly bound between two acidic residues in a similar fashion expected for a xylose substrate/product, as also happened with the *G. thermoglucosidasius* β-xylosidase-1 (4.3.6.1). These two carboxylic amino acid were later identified by the authors as the catalytic residues, Glu134 (the acid/base) and Glu241 (the nucleophile), and as in XT6 they were located near the C-terminal ends of β-strands 4 and 7, respectively, and separated by 5.54 Å. The work of Zolotnitsky *et al.* (2004) also helped to elucidate the xylosaccharides binding thermodynamics of IXT6 and Figure 5.12 shows its active site and the amino acids responsible for the catalysis, substrate binding and stacking interactions.

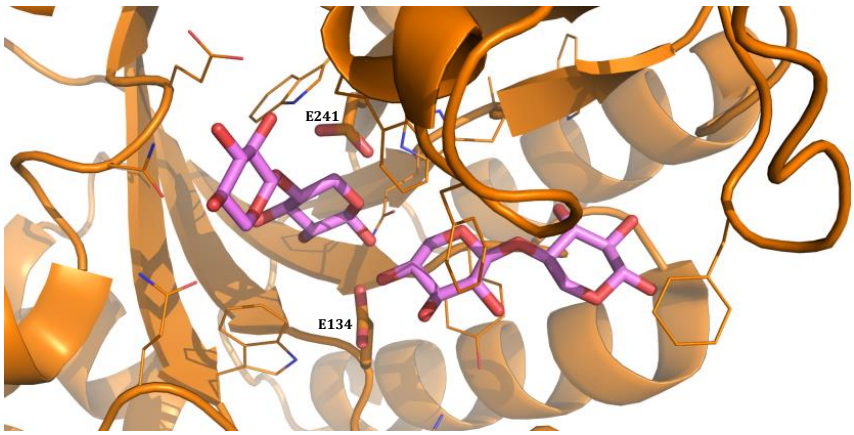


Figure 5.12. IXT6 active-site cleft in complex with cleaved xylotetraose (PDB entry: 3MUI). Residues forming the cleft are shown in stick form and the carboxylic catalytic residues D134 and E241 are shown in thicker lines.

Even though from Figure 5.11 it can be seen that the substrates of both xylanases are bound in the same place, a closer view of the substrate-binding clefts of IXT6 and XT6 shows significant differences in their overall shape and dimensions. The cleft in IXT6 is produced by several side chains that extend into the inner part of the cleft, as well as several sections of the polypeptide main chain that are displaced inwards into the cleft. As a result, the IXT6 active site shown in Figure 5.13 is shorter, narrower, and nearly blocked at the two ends, and its walls seem to be nearly perpendicular to the bottom of the cleft. On the other hand, as previously shown in Figure 5.7, the corresponding cleft in XT6 is much longer, wider, and open at both ends, and its walls seem to form a wider angle facing the outside solvent area. Furthermore, ITC data and the overlap of both xylanase crystal structures indicate that the IXT6 active site consists of only five binding subsites (designated accordingly from -3 to +2), compared with the six subsites of XT6, and the specific amino acid residues associated with the particular substrate-binding subsites are: Gln88(-3), Trp291(-2), Trp299(-1), Tyr179(+1) and Phe249(+2) (Solomon *et al.*, 2007; Zolotnitsky *et al.*, 2004).

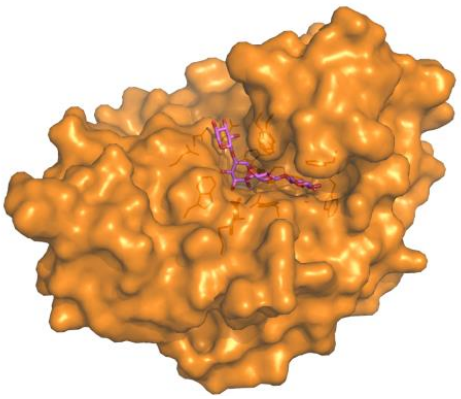


Figure 5.13. IXT6 active-site groove in complex with its xylotetraose substrate shown in stick form (PDB entry: 3MUI). A surface representation with 20% transparency.

A structure-based sequence alignment of the extracellular and intracellular xylanases from *Geobacillus stearothermophilus* T-6, C56YS93, CAMR12739 and CAMR5420 is shown in Figure 5.14, and it can be seen that the catalytic residues and the amino acids associated with the substrate-binding subsites are almost fully conserved in both the extracellular and intracellular enzymes from the four different *Geobacillus* strains; the only exception is the residue associated with the +3 subsite in the extracellular xylanases that is not present in the intracellular ones. These comparisons indicate that the crystal structures deposited in the PDB provide enough comparable information for structurally-based engineering of the xylanases from C56-YS93, CAMR12739 and CAMR5420 microorganisms.

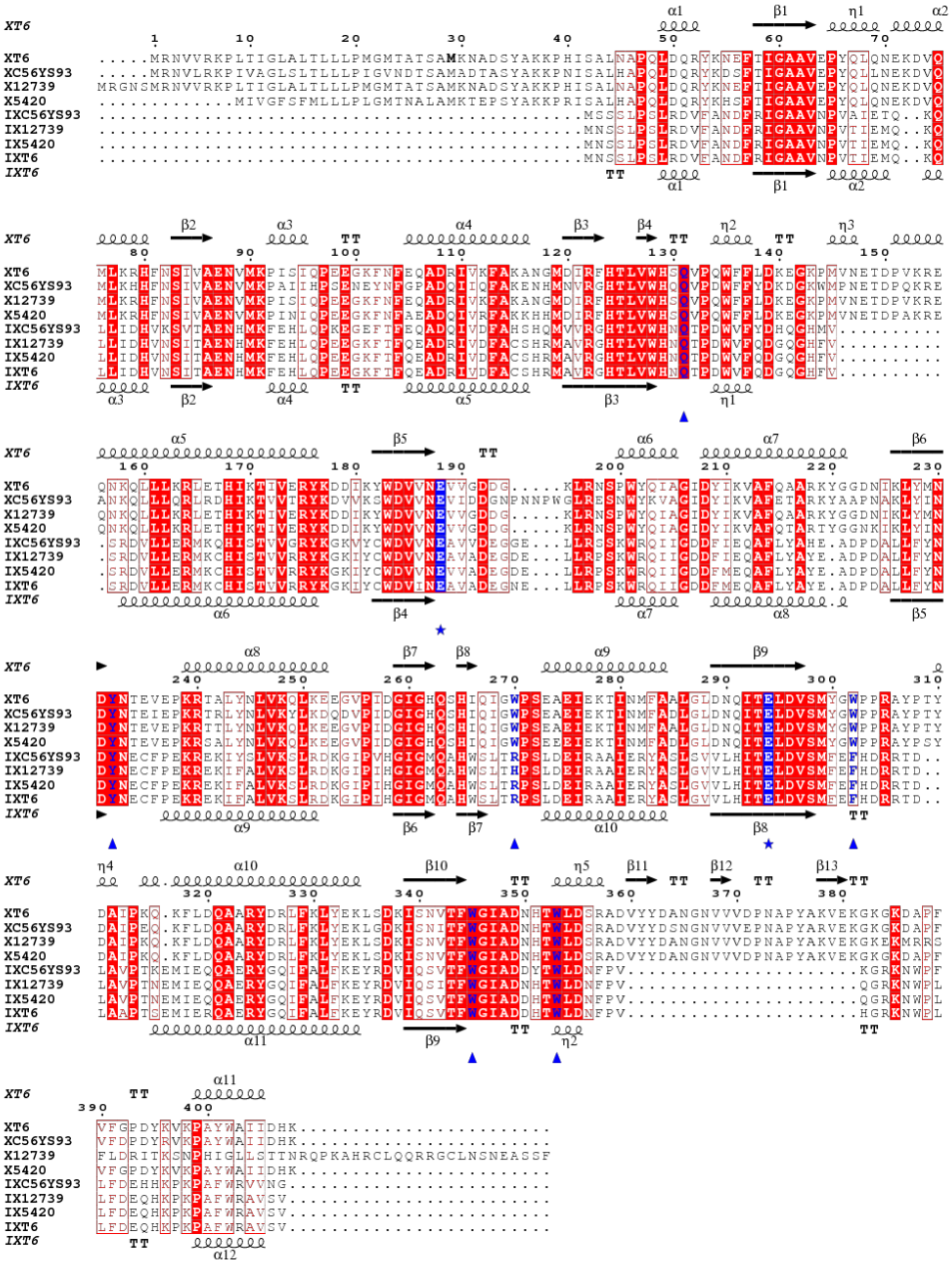


Figure 5.14. Structure-based sequence alignment of GH10 extracellular (X) and intracellular (IX) xylanases from *G. stearothermophilus* T-6, *G. thermoglucosidasius* C56-YS93, CAMR 12739 (most likely *G. stearothermophilus/G. thermoleovorans)* and CAMR 5420 (most likely *G. kaustophilus)* produced using Clustal Omega and displayed using ESPript 3.06 (Gouet *et al.*, 1999). Identical residues are shown in white on a red background, while conservative amino acid changes are shown in red. Amino acids involved in the interaction with substrate are highlighted in blue (the catalytic residues are indicated by blue stars, and the others, associated with the substrate binding subsites, by blue triangles beneath the alignment). Secondary structure of *G. stearothermophilus* T-6 extracellular xylanase, XT6, is shown above the alignment, while the secondary structure of its intracellular xylanase, IXT6, is shown beneath the alignment with β-strands as arrows and helices as coils. It should be noted that the predicted signal peptide sequences are included for comparison, but the initial methionine of the mature protein would begin at the residue 29 (in bold) in this alignment.

5.3.4. Xylanolytic activity of the recombinant xylanases expressed in E.coli

5.3.4.1. Xylanases from G. thermoglucosidasius C56-YS93

Three different sources of xylan, beechwood, birchwood and oat spelt, were used to assess the xylanolytic activity of the purified recombinant xylanases, XC56 and IXC56, from *G. thermoglucosidasius* C56-YS93. A negative control was carried out using the recombinant TM242 alcohol dehydrogenase (kindly donated by Dr Jonathan Extance, University of Bath) that does not possess activity against xylan. The activity of the recombinant TM242 β-xylosidase-1 (described in Section 3.3.3.1 and Chapter 4) against xylan was also evaluated and the results are shown in Figure 5.15.

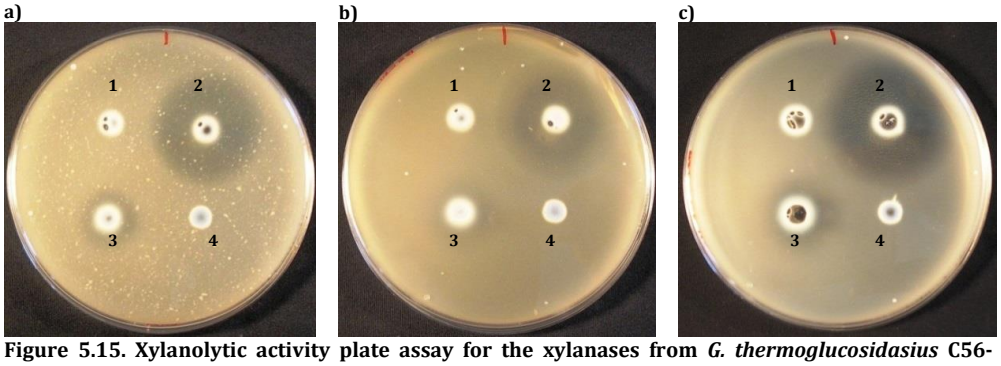


Figure 5.15. Xylanolytic activity plate assay for the xylanases from *G. thermoglucosidasius* C56-YS93. USM agar plates containing: (a) 1% (w/v) oat spelt xylan, (b) 1% (w/v) birchwood xylan, and (c) 1% (w/v) beechwood xylan. The wells correspond to (1) β-xylosidase-1, (2) XC56, (3) IXC56, (4) Negative control: Alcohol dehydrogenase; each well contains 40 µL purified protein at a concentration of ~2 mg/mL.

As can be seen from Figure 5.15, both recombinant xylanases produced a halo indicating xylan degradation, while both the negative control and β-xylosidase-1 do not seem to hydrolyse any of the substrates tested. As discussed in Chapter 4 (section 4.3.9.2 and 4.3.10), the crystal structure of β-xylosidase-1 revealed a restricted active site accessibility, limiting the access to smaller xylo-oligosaccharides rather than bulk polymers. This suggests no activity would be found with xylan as it cannot fit in the active site pocket (Espina *et al.*, 2014). Figure 5.15 also shows that the clearing zone of XC56 is significantly larger than IXC56 indicating the extracellular enzyme (cloned without its signal peptide) possesses a superior xylanolytic activity than the intracellular xylanase. The protein concentration added to each well was very similar (~2 mg/mL), indicating that XC56 has a higher specific activity than IXC56 against xylan (from the three different sources tested). In addition, this xylanolytic behaviour is very similar in the three different sources of xylan tested, where it is possible to see the same clearing tendency, with the highest xylanolytic activity being observed in plates containing beechwood xylan, followed by oat spelt.

The xylanolytic activity was also measured by means of the DNS assay, which can be considered as a semi-quantitative assay because the enzyme activity is calculated as xylose equivalents (from a calibration curve) even though most xylanases produce little xylose and mostly release reducing sugar oligomers (*e.g.* xylobiose and xylotriose) that have been described to be more reactive with DNS than xylose (Jeffries *et al.*, 1998), leading to over-estimations of the xylanolytic activity (Gusakov *et al.*, 2011). Despite its limitations, the DNS assay is widely used in the literature as it allows rapid measurement of reducing ends, which makes it suitable for comparing the relative enzymatic activities of different samples. In the case of the recombinant xylanases from *G. thermoglucosidasius* C56-YS93, various volumes of each enzyme were incubated with 1% (w/v) birchwood xylan, at 60°C, with a one-minute assay giving a linear rate (Figure 5.16). Although these results were not optimised (or repeated) due to time constraints, the DNS colour change confirmed the xylanolytic activity of XC56 to be much higher than IXC56, as was observed on the plate assays.

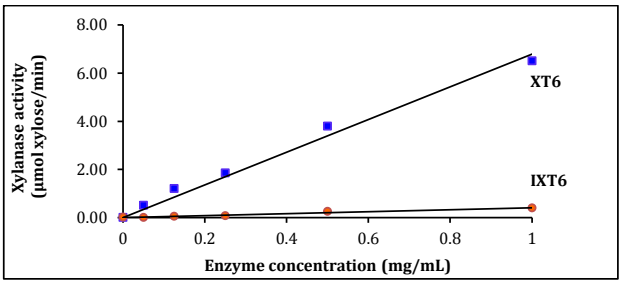


Figure 5.16. DNS assay of the recombinant xylanases from *G. thermoglucosidasius* C56-YS93. Various volumes of each enzyme were incubated with 1% (w/v) birchwood xylan for one minute at 60°C, prior to the addition of DNS, and the absorbance of the coloured complex was read at 540 nm. The xylanolytic activity of the enzymes was calculated as the amount of reducing sugars that react with the organic DNS compound.

On the other hand, continuous spectrophotometric assays using artificial substrates (performed as described in 3.2.8) revealed that IXC56 possesses high hydrolytic activity against *p*NP-XP (with a specific activity of 30 U mg^{-1}) and is even able to hydrolyse *p*NP-AP (with a specific activity of 5 U mg^{-1}), while no (or extremely low) activity against these substrates was observed with XC56.

These observations indicate that the intracellular xylanase, IXT6, from *G. thermoglucosidasius* C56-YS93 behaves more similarly to a β-xylosidase than a xylanase, but due to its ability to degrade small amounts of xylan, has been annotated as a xylanase instead of a true β-xylosidase. The lack of a signal peptide sequence also

indicates that this enzyme is more specialised in the intracellular degradation of xylo-oligosaccharides to xylose rather than xylan, as it is unlikely that the decorated xylan polymer can enter the cell to be degraded by an intracellular xylanase.

The results obtained are consistent with the structural differences discussed above (and observed in Figure 5.7 and Figure 5.13), where the crystal structures of extracellular xylanases show an open, longer and wider active site pocket, which allows the binding of the long and branched xylan substrates. On the other hand, the structure of intracellular xylanases revealed a blocked, shorter and narrower active site cleft that seems to prefer short and linear xylo-oligosaccharides substrates (Solomon *et al.*, 2007).

In the case of the most extensively studied xylanolytic *Geobacillus* strain, *G. stearothermophilus* T-6, the substrate specificity towards xylan and its fragments has also been proven to be significantly different for its two xylanases (Zolotnitsky *et al.*, 2004). Fortunately, the molecular determinants for hemicellulose degradation and utilisation by this microorganism have been identified and partially characterised (Shulami *et al.*, 2014).

It has been described that the utilisation of plant cell wall biomass by *G. stearothermophilus* T-6 is based on several elements (summarised in Figure 5.17). At first, it involves a two-component extracellular oligosaccharide-sensing system XynDC, that detects the presence of xylan in the environment by sensing small amounts of xylose, which presumably exist as a by product from the hydrolytic activity of neighboring microorganisms. The xylose interacts with the extracellular domain of the class I histidine kinase sensor protein, XynD, and triggers the phosphorylation of the response regulator XynC. Then, the phosphorylated XynC activates the expression of a dedicated ATP-binding cassette (ABC) sugar transporter (XynEFG) that facilitates the entrance of xylo-oligosaccharides into the cell. Their subsequent hydrolysis to xylose allows the induction of the xylan-utilization genes via inactivation of the XylR repressor (Shulami *et al.*, 2007; 2014).

The efficient xylan utilisation strategy is then initiated by the expression of a single extracellular xylanase (XT6), which is not only negatively controlled by the xylose repressor, XylR, and induced by xylose, but is also repressed by glucose via Carbon Catabolite Repression, where the XT6 gene is repressed in media that contain glucose

and casamino acids. In addition, the gene is regulated by two yet uncharacterised mechanisms, a NIF3-like protein, XynX, and by cell density (quorum sensing) (Shulami *et al.*, 2014).

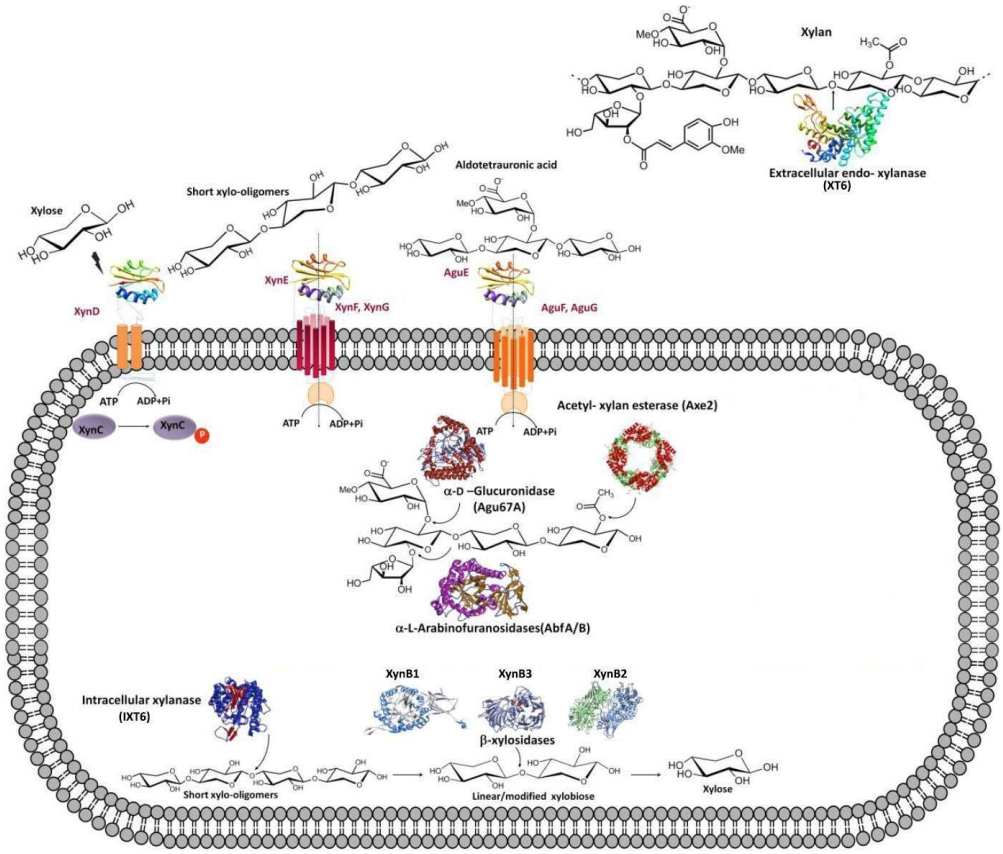


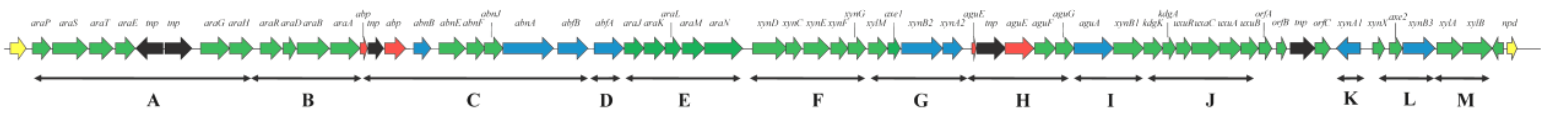
Figure 5.17. Schematic view of the xylanolytic system from *G. stearothermophilus* T-6. The utilisation of xylan by the microorganism is initiated by the presence of extracellular xylose that interacts with the extracellular domain of XynD, which then phosphorylates XynC. The phosphorylated XynC activates the expression of a dedicated ABC xylo-oligosaccharides transporter, XynEFG, which facilitates the entrance of xylo-oligosaccharides into the cell, serving as a molecular inducer. Expression of the extracellular xylanase (XT6) provides the bacterium with decorated xylo-oligosaccharides that enter the cell via the ABC transporter XynEFG and AguEFG (for xylo-oligosaccharides and aldotetrauronic acid, respectively). The decorated xylo-oligomers are then hydrolysed to their corresponding monomers by intracellular side-chain-cleaving enzymes, including α-glucuronidase (Agu67A), two GH51 α-arabinofuranosidases (AbfA and AbfB), two xylan acetyl esterases (CE4), the intracellular xylanase (IXT6) and three β-xylosidases (XynB1 (GH39), XynB2 (GH52), and XynB3 (GH43)). Image modified from Shulami *et al.* 2014.

The secretion of XT6 initiates the hydrolysis of the long branched xylan polymer backbone, producing shorter substituted oligosaccharide products decorated with various side-chains such as L-arabinose, 4-O-methyl glucuronic acid and acetate. These are subsequently carried across the cellular membrane by specialised ABC sugar transporters (XynEFG for xylo-oligosaccharides and AguEFG for aldotetrauronic acid) that allow their uptake (Shulami *et al.*, 2007; 2014). Inside the cell, the side

substitutions are removed by the action of intracellular debranching enzymes, (including two GH51 α-arabinofuranosidases (AbfA and AbfB), a GH67 α-glucuronidase (Agu67A), and two CE4 acetyl xylan esterases (Axe2)) resulting in short unmodified (linear) xylo-oligosaccharides. These are further cleaved into D-xylose monomers through the concerted action of the intracellular xylanase (IXT6) and three different β-xylosidases (XynB1 (GH39), XynB2 (GH52), and XynB3 (GH43)), the products of which are then metabolised via the pentose phosphate pathway and glycolysis (as previously shown in Figure 1.7) (Shulami *et al.*, 1999; 2007; 2014).

It has been shown that the ability of *G. stearothermophilus* T-6 to degrade and utilise hemicellulose is dependent on a single complex gene locus, which is ~78 kb in size, has a G + C content of 47%, encodes 60 proteins and is localised on a common genomic island (De Maayer *et al.*, 2014; Shulami *et al.*, 2011). This Hemicellulose Utilisation System (HUS) locus (Figure 5.18) can be roughly divided into thirteen distinct gene clusters (A-M) involved in the concerted degradation and utilisation of the hemicellulose polymer glucuronoarabinoxylan (De Maayer *et al.*, 2014).

To date, the genomes of twenty-five *Geobacillus* spp. have been sequenced and, according to the bioinformatic analysis performed by De Maayer *et al.* (2014), eighteen strains (72%) can be considered to be HUS+. The comparison of the different *Geobacillus* orthologous hemicellulose utilisation loci reveals extensive variation among the strains, which will explain why some of the main enzymes related to hemicellulose hydrolysis encoded in cluster K are not present in *G. thermoglucosidasius* TM242 (even though this strain is very similar to *G. thermoglucosidasius* C56-YS93 and shares 100% identity in central metabolic enzymes including glucokinase (ZP_06809151.1), citrate synthase (ZP_06809481.1) and ribose 5-phosphate isomerase (ZP_06810657.1). This genetic variability among HUS loci implies that they have variable capacities to degrade hemicellulose polymers (probably depending on their environmental exposure to a particular source), or that they may degrade distinct polymers, as are found in different plant species and tissues. Consequently,diverse species living in symbiotic communities may complement each other, and this could potentially be exploited to generate recombinant *Geobacillus* strains with improved hemicellulolytic capacities and/or with the ability to degrade structurally diverse hemicellulose substrates (De Maayer *et al.,* 2014).



Cluster	Function
A	L-Arabinose transport
B	Arabinose metabolism
C	Arabinan utilization
D	Intracellular arabinofuranose degradation
E	Alternative pathway for pentose sugar metabolism

Cluster	
F	Xylooligosaccharide transport
G	Intracellular xylooligosaccharide degradation
H	Aldotetraouronic acid transport
I	Intracellular Aadotetraouronic acid degradation
J	Glucuronic acid metabolism

Cluster	
K	Extracellular xylan degradation
L	Intracellular xylooligosaccharide degradation
M	Xylose metabolism

Figure 5.18. Schematic diagram of the *G. stearothermophilus* T-6 hemicellulose utilisation locus subdivided into thirteen gene clusters on the basis of their predicted function. Each arrow represents a gene in the locus, with genes encoding predicted transposons coloured in black, while open reading frames interrupted by transposons are coloured in red. Genes encoding glycosyl hydrolases are coloured in blue. Image obtained from De Maayer *et al.,* 2014a.

5.3.4.2. Xylanases from Geobacillus spp. CAMR12739 and CAMR5420

Beechwood xylan was used to assess xylanolytic activity of the purified recombinant xylanases, X12739 from CAMR12739 (probably *G. stearothermophilus* or *G. thermoleovorans),* and X5420 from CAMR5420 (most likely *G. kaustophilus*). Consistent with the previous experiment (Figure 5.15), the recombinant β-xylosidase-1 was used as a negative control, while XC56 was used as a positive control. The results are shown in Figure 5.19.

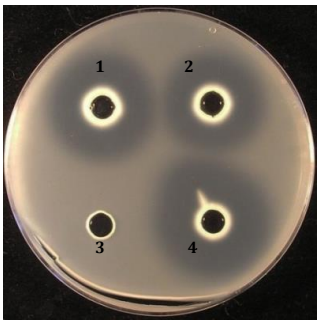


Figure 5.19. Xylanolytic activity plate assay for the xylanases from *Geobacillus* spp. CAMR12739 and CAMR5420. USM agar plates contained 1% (w/v) beechwood xylan with the wells corresponding to (1) X12739, (2) X5420, (3) Negative control: β-xylosidase-1, (4) Positive control: XC56; each well contains 40 µL purified protein at a concentration of ~2 mg/mL.

As can be seen from Figure 5.19, X12739, X5420 and the positive control produce a halo indicating xylan degradation, while the negative control, as expected, did not hydrolyse the substrate. The clearing zone of X12739 is bigger than that with X5420 suggesting the recombinant enzyme from CAMR12739 possesses a superior xylanolytic activity than the one from CAMR5420. As the protein concentration added to each well was very similar (~2 mg/mL), it is possible to infer that X12739 has a higher specific activity than X5420 against beechwood xylan; however, it is not superior to the xylanolytic activity of the positive control, XC56 from *G. thermoglucosidasius* C56-YS93.

The hydrolytic activity against birchwood and oat spelt xylans could not be assessed because the manufacturer (Sigma-Aldrich) discontinued the xylan from these other sources, and at the time of the experiments xylan from beechwood was the only substrate available.

The DNS assays confirmed the same results obtained with the plate assays (Figure 5.20), with X12739 being more active than X5420, but less active than XC56.

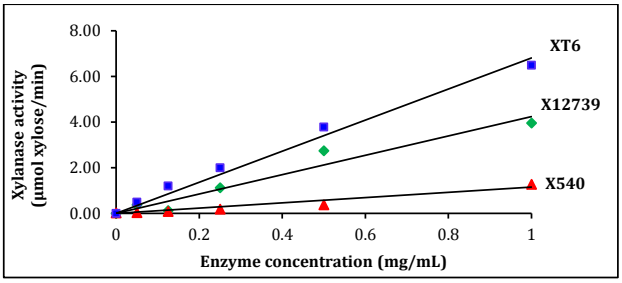


Figure 5.20. DNS assay of the recombinant xylanases from CAMR12739 and CAMR5420. Various volumes of each enzyme were incubated with 1% (w/v) birchwood xylan for one minute at 60°C, prior to the addition of DNS, and the absorbance of the coloured complex was read at 540 nm. The xylanolytic activity of the enzymes was calculated as the amount of reducing sugars that react with the organic DNS compound.

5.3.5. Amplification, cloning and heterologous expression in TM242

The three extracellular xylanase-encoding genes (one from each strain) were selected for further study: cloning and recombinant expression in *G. thermoglucosidasius* TM242. The three xylanase-encoding genes, including their predicted signal peptide sequences, were amplified using their respective *Geobacillus* genomic DNAs as template. Specific PCR products of the approximate expected size were obtained as shown in Figure 5.21. The bands were excised from the gel, A-tailed and cloned into pGEM®-T Easy, and then subcloned into the *Geobacillus* spp. expression vector pUCG4.8.

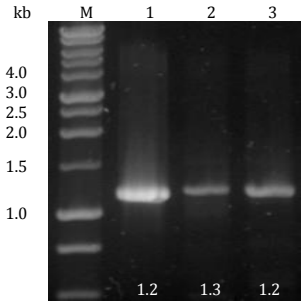


Figure 5.21. PCR amplification of xylanase-encoding genes from different *Geobacillus* genomic DNA. Amplicons are shown on a 0.8% agarose gel corresponding to: (1) XC56 extracellular xylanase from *G. thermoglucosidasius* C56-YS93 (including its signal peptide), (2) X12739 extracellular xylanase from CAMR12739 (including its signal peptide), (3) X5420 extracellular xylanase from CAMR5420 (including its signal peptide), and (M) marker lanes containing a 1 kb ladder (Promega). The expected sizes in kb for each PCR product are shown below each band.

Sequencing confirmed the cloned gene sequences were correct and the vectors harbouring the xylanase-encoding genes were then transformed into *G. thermoglucosidasius* TM242 and grown in 2SPYNG medium. Protein expression using pUCG4.8 vector is under the control of the uracil constituve promoter PUp2n38, and therefore no induction or specific growth conditions were required. After 6 h the cells were harvested and the supernatants were collected and concentrated in order to analyse the expression and secretion of the heterologous xylanases.

The successful expression and secretion of the heterologous XC56 in TM242 was confirmed by Western Blotting (Figure 5.22) using specific antibodies raised by Eurogentec (Southampton, UK) against purified recombinant XC56 expressed in *E. coli*. This experiment was performed by the University of Bath PhD candidate Alexandria Holland, and the results are produced here with her permission.

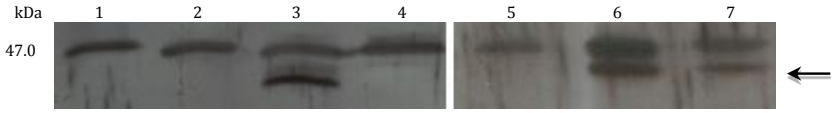


Figure 5.22. Western Blot analysis of the expression and secretion of XC56. (1) *G. thermoglucosidasius* C56-YS93 cell pellet, (2) TM242 cell extract, (3) C56 concentrated supernatant, (4) TM242 concentrated supernatant, (5) TM242 concentrated supernatant, (6) concentrated supernatant from TM242 harbouring pUCG4.8-XC56, (7) C56 concentrated supernatant. The extracellular xylanase from C56 is highlighted with an arrow (predicted M_r = 44.6 kDa).

Unfortunately, the heterologous expression and secretion of the other two extracellular xylanases, X12739 and X5420, was not achieved in the time available, even though several attempts were made using the different vectors and promoters described by Bartosiak-Jentys *et al.* (2013). However, as previously shown in Figures 5.19 and 5.20, XC56 appears to be catalytically superior to the other two xylanases when expressed in *E. coli,* suggesting that even successful expression and secretion of X12739 and X5420 in *G. thermoglucosidasius* TM242 might not be better than the results obtained with XC56.

5.3.6. Xylanolytic activity of the heterologous extracellular xylanase from G. thermoglucosidasius C56-YS93 expressed in TM242.

Plate assays, employing two different sources of xylan, beechwood and oat spelt, were used to assess and compare the xylanolytic activity of *G. thermoglucosidasius* TM242 harbouring the expression vector pUCG4.8 with the gene encoding the extracellular xylanase from *G. thermoglucosidasius* C56-YS93.

Successful expression and secretion of the heterologous extracellular xylanase from *G. thermoglucosidasius* C56-YS93 can be observed in Figure 5.23, as the colony of the cells harbouring the pUCG4.8-XC56 plasmid as well as its supernatant produced a halo indicating xylan degradation, while no expression is observed in the untransformed cells.

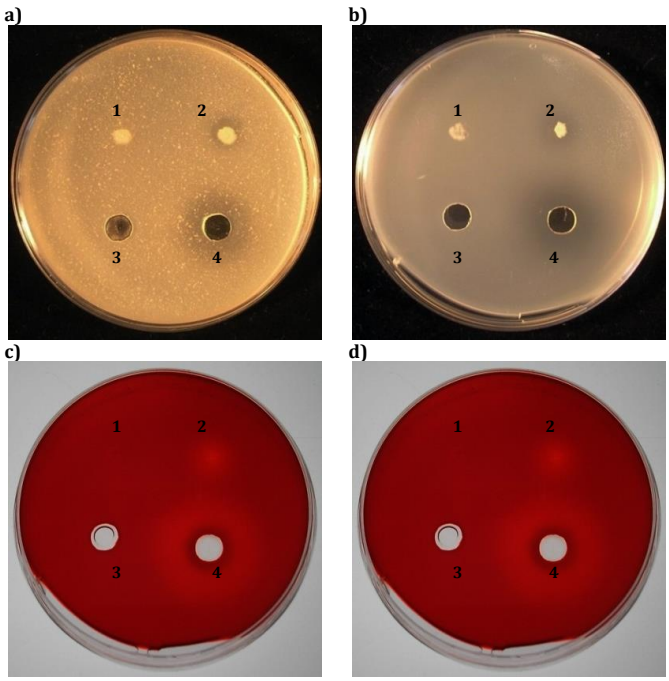


Figure 5.23. Xylanolytic activity plate assay for the extracellular xylanase from *G. thermoglucosidasius* C56-YS93 heterologously expressed in TM242. USM agar plates contained (a) 1% (w/v) oat spelt xylan, (b) 1% beechwoodxylan. The wells correspond to (1) *G. thermoglucosidasius* TM242 colony, (2) *G. thermoglucosidasius* TM242 harbouring pUCG4.8-XC56 colony, (3) 40 µL of concentrated supernatant from TM242, (4) 40 µL of concentrated supernatant from TM242 harbouring pUCG4.8-XC56. Panels (c) and (d) are plates (a) and (b), respectively, stained with 0.1% (w/v) Congo Red aqueous solution as described in Section 5.2.8.1.

Although quantitative analysis of the heterologous XC56 xylanolytic activity was not possible in the timescale of this project, further characterisation of this enzyme is currently under investigation by the PhD candidate Alexandria Holland, including purification of the enzyme, assays of its xylanolytic activity in a quantitative manner to determine its kinetic parameters, and an analysis of the mechanism and control of its secretion.

5.4. CONCLUSIONS

Four different xylanase-encoding genes from closely related thermophilic microorganisms (two from *G. thermoglucosidasius* C56- YS93, one from *G. kaustophilus* and one from *G. stearothermophilus/G. thermoleovorans*) have been successfully cloned and expressed in *E. coli* BL21 using the pET-28a(+), and the soluble recombinant proteins have been purified by nickel-affinity chromatography.

The recombinant enzymes were shown to be active using two discontinuous assay methods: production of clearing zones on plates of minimal media containing xylan, and detection of xylan hydrolysis using DNS. The extracellular xylanase from *G. thermoglucosidasius* C56-YS93 (XC56) showed the highest xylanolytic activity, followed by the extracellular xylanases from CAMR 12739 (X12739) and CAMR5420 (X5420); all enzymes were recombinantly expressed without their signal peptides. As expected, the IXC56 xylanase from *G. thermoglucosidasius* C56-YS93 showed very little activity against xylan, but was highly active against *p*NP-β-D-xylopyranoside, confirming it behaves more as a β-xylosidase than a proper xylanase, a result that is consistent with its intracellular location.

The genes encoding the extracellular xylanases (XC56, X12739 and X5420) were also cloned with their corresponding signal peptide sequences in an expression vector for heterologous constitutive expression in *Geobacillus.* Successful expression in *G. thermoglucosidasius* TM242 and secretion of the heterologous extracellular xylanase XC56 was achieved. This enzyme was shown to display enzymatic activity in both qualitative discontinuous assays tested.

These results are promising and it is felt that have the potential to make a valuable contribution towards second-generation bioethanol production, as the heterologous expression and secretion have the potential to effect enhanced biomass utilisation by TM242, contributing towards the development of consolidated bioprocessing of lignocellulose to bioethanol.

CHAPTER 6: Conclusions and Future Work

As described in Chapter 1, the organism *G. thermoglucosidasius* TM242 has the potential to offer commercially-viable bioethanol production on a large scale. It utilises a wide range of sustainable feedstocks (efficiently fermenting both C6 and C5 sugars), and it has been genetically engineered to divert the fermentation pathways away from mixed acids to generate ethanol as the major product. However, hydrolysis of biomass feedstock with commercial enzymes remains a major economic consideration in the production of second-generation bioethanol as an alternative fuel. For this reason, the work presented throughout this thesis aims to improve this cost-limiting step of bioethanol production and enhance biomass utilisation through an investigation of the enzymes mainly associated with hemicellulose hydrolysis.

6.1. Recombinant glycosidases from *G. thermoglucosidasius* TM242

The identification of a number of genes encoding glycoside-hydrolases in the TM242 genome sequence has been described in Chapter 3. The selective cloning of six of them, and the successful expression of the recombinant proteins in *E. coli*, gave three thermostable carbohydrate-degrading enzymes associated with hemicellulose hydrolysis: two β-xylosidases (one from GH52 and the other from GH43) and an α-arabinofuranosidase (from GH51), all of which are thought to play a key role during sugar metabolism in the TM242 microorganism, and may be important for bioethanol production.

These recombinant hemicellulases were found to be highly active (using *p*NP-linked glycosides as substrates) and were characterised in terms of their temperature optimum, thermostability, substrate specificity and kinetic parameters. The three recombinant enzymes had optimal temperatures expected for the thermophilic microorganism *G. thermoglucosidasius* TM242 and, even though expressed in a mesophilic host, they were relatively stable between temperatures of 60-70°C over a 60-minute period. These results were comparable with the ones in the literature for the same GH family enzymes from closely related thermophilic microorganisms

Michaelis-Menten kinetics were observed with all the substrates tested for each enzyme. In the case of β–xylosidase-1, the results obtained were very similar to the kinetics of the GH52 β–xylosidase enzyme XynB2 from *G. stearothermophilus* T-6

(Bravman *et al.*, 2003), although the k_{cat} was far superior for β-xylosidase-1 (at least 3-fold higher) with all the artificial substrates tested. Furthermore, the k_{cat} values were also superior for β-xylosidase-1 than for the recombinant TM242 β-xylosidase-2 (45 times higher with *p*NP-XP and 2.5 higher with *p*NP-AF), which appears to catalyse the hydrolysis of the artificial substrates with much lower catalytic efficiency in comparison with β-xylosidase-1 (17-fold lower with *p*NP-XP and 3.5 fold lower with *p*NP-AF). On the other hand, the data obtained with the recombinant α-arabinofuranosidase were comparable to the kinetics of the GH51 α-arabinofuranoside enzyme AbfA also from *G. stearothermophilus* T-6 (Shallom *et al.*, 2002), with superior catalytic efficiency and k_{cat} values for the TM242 α-arabinofuranosidase over AbfA (2-fold and 3-fold higher, respectively, with *p*NP-AF).

Further characterisation of the enzymes with their natural substrates was carried out and the results obtained indicate that β-xylosidase-1 is able to hydrolyse short xylo-oligosaccharides (xylobiose and xylotriose) while no activity with xylan was found; these same results were also observed (qualitatively) in the case of β-xylosidase-2 (Cronin, 2013). α-Arabinofuranosidase was shown to be able to hydrolyse arabinan (Goffin, 2014) and the catalytic activity with other arabinose-containing polymers such as sugar beet pulp is currently under investigation by Dr Charlotte Bennett (University of Bath).

According to a structural search of the PDB with the aminoacid sequences of these three enzymes, β-xylosidase-2 and α-arabinofuranosidase share the highest sequence similarities (85-90%) with those of their respective published structures from *G. stearothermophilus* T-6 (obtained at 1.88 Å and 2 Å resolution, respectively). Furthermore, the amino acid residues described to be crucial for both enzymic activities were shown to be fully conserved in the TM242 enzymes, indicating that the crystal structures deposited in the PDB provide comparable enough information for structurally-based engineering of the TM242 enzymes. However, at the time of the studies done with β-xylosidase-1, there was no crystal structure of this protein or of any member of the family GH52. For this reason, selenomethionine derivatised crystals were grown and high-resolution X-ray data were obtained at PETRA III synchrotron (EMBL). After processing, phasing, model building and refinement, a high-resolution (1.7 Å) crystal structure of the apo-enzyme along with a lower resolution (2.6 Å) structure of the enzyme-substrate complex, were determined during this project, resulting in the first reported structure of a GH52 family member (Espina *et al.*, 2014).

The novel β-xylosidase structure reveals that the GH52 enzymes are composed of two domains, an N-terminal β-sandwich and a C-terminal $(\alpha/\alpha)_6$ barrel. Surprisingly, the fold of this enzyme shares structural similarities with other glycoside-hydrolases, despite sharing less than 13% sequence identity, whilst it is completely different from the xylosidase structures in other CAZy families. The active site of the enzyme is located at the centre of the $(\alpha/\alpha)_6$ domain and the positioning of the xylobiose substrate is consistent with the family's proposed retaining mechanism; additionally, the deep cleft of the active site pocket, plus the proximity of the neighbouring subunit, afford an explanation for the lack of catalytic activity towards the polymer xylan.

This first GH52 structure is important as it offers a valuable resource for anyone working with a family GH52 enzyme since the coordinates of the apo-enzyme and enzyme-substrate complex have been deposited in the PDB and are now publicly available. For this particular project, this novel structure is relevant as it provides the basis for structurally-informed engineering of the enzyme for improved xylan degradation and bioethanol production.

The potential for rational design to engineer the specificity of β-xylosidases and the α-arabinofuranosidases have been demonstrated on several occasions where, through site-directed mutagenesis, the enzymes are able to display new catalytic activities (*e.g.* xylanase) while retaining their original activities (McKee *et al.,* 2012; Huang *et al.,* 2014). It has also been suggested that these enzymes might provide a structural platform that can be harnessed to bind a range of different sugars and to catalyse the hydrolysis of glycosidic bonds through distinct modes of action in order to generate multifunctional enzymes (McKee *et al.*, 2012; Huang *et al.,* 2014), which would be particularly useful during bioethanol production.

6.2. Recombinant xylanases from closely related *Geobacillus*

The search of the TM242 genome sequence revealed this microorganism lacks genes encoding xylanases, the main enzymes required for thorough hemicellulose degradation. The wild-type microorganism, *G. thermoglucosidasius* NCIMB 11955 was isolated from a compost heap environment which would contain a wide variety of other thermophilic strains forming a consortium where strains deficient in genes or pathways required for hemicellulose degradation and utilisation may be complemented by other members of the consortium. Therefore, it can be hypothesised

that 11955, in its ability to utilise a wide range of sugars, just used the oligosaccharides that other microoorganisms produced with no need to express a xylanase.

On the other hand, it has been shown that closely related *Geobacillus* strains possess at least one xylanase-encoding gene in their genome (Anand *et al.*, 2013; Canakci *et al.*, 2012; Khasin *et al.*, 1993; Liu *et al.*, 2012). De Maayer *et al.* (2014) suggest that the highly variable hemicellulolytic capacities of different *Geobacillus* strains imply that they have evolved to degrade distinct hemicellulose substrates (such as are found in different plant species and tissues); furthermore, their ability to degrade distinct hemicellulose polymers may thus reflect their environmental exposure to particular hemicelluloses and, potentially, their role in carbon cycling in places where hot conditions persist (De Maayer *et al.*, 2014a).

G. stearothermophilus T-6 is the most studied xylanolytic *Geobacillus* strain and its molecular determinants for hemicellulose degradation and utilisation have been identified and partially characterised (Shulami *et al.*, 2014). In this strain, the initial hydrolysis of the long branched xylan polymer is executed by an extracellular xylanase that degrades the polysaccharide backbone into short substituted xylo-oligosaccharides, which can then enter into the cell by virtue of specific ABC sugar transporters (Shulami *et al.*, 2007, 2011). Inside the cell, the side substitutions are removed by the action of intracellular debranching enzymes, resulting in short unmodified (linear) xylo-oligosaccharides that are then hydrolysed by an intracellular xylanase and different β-xylosidases (Solomon *et al.,* 2007).

This physiological strategy, limiting the number of secreted hydrolytic enzymes, has been proposed to provide a selective advantage to *G. stearothermophilus* T-6 and related hemicellulolytic thermophiles (Shulami *et al.*, 2011). It allows the bacteria to successfully compete in nature with other plant cell-wall degrading microorganisms as the relatively large oligosaccharides (which result from extracellular hemicellulose hydrolysis) are not readily available to competing, non-hemicellulolytic, microorganisms because they lack the appropriate sugar transporters or intracellular enzymes required to breakdown short substituted xylo-oligosaccharides (which is not the case in *G. thermoglucosidasius* 11955).

The cloning of genes encoding xylanases from closely-related *Geobacillus* strains has been described in Chapter 5. The strains available during this project were *G. thermoglucosidasius* C56-YS93, CAMR12739 (most likely *G. stearothermophilus* or

G. thermoleovorans) and CAMR5420 (most likely *G. kaustophilus*). According to their genomes, each strain possesses two xylanase-encoding genes, one with a signal peptide and therefore encoding an extracellular xylanase, and one without a signal peptide that presumably encodes an intracellular xylanase, just as described for *G. stearothermophilus* T-6.

According to a structural search of the PDB with the aminoacid sequences of these six enzymes, the intracellular and extracellular xylanases share the highest sequence similarities (88-99%) with the published structures from *G. stearothermophilus* T-6 (obtained at 2.4 Å and 1.45 Å resolution, respectively). Furthermore, the catalytic residues and the amino acids associated with the substrate-binding subsites are almost fully conserved in both the extracellular and intracellular enzymes from the four different *Geobacillus* strains. These results indicate that the crystal structures deposited in the PDB provide enough comparable information for structurally-based engineering of these enzymes, which is very promising as the potential for improvement of the thermostability and catalytic efficiency of xylanases from *Geobacillus* spp has been demonstrated before. Zhang *et al.* (2010) enhanced by 52-fold the half-life of inactivation of XT6 at 75°C and increased by 90% its catalytic efficiency through directed evolution in combination with consensus-based semi-rational approach and site-directed mutagenesis while Wang *et al.* (2013), used error prone PCR and site-saturation mutagenesis to increase 3.2-fold the catalytic efficiency of the intracellular xylanase from *G. stearothermophilus* 1A05583 (85% identical to IXT6) mutant H179F .

The genes selected for cloning were the extracellular (XC56) and intracellular (IXC56) xylanases from C56-YS93 and the extracellular xylanases from CAMR 12739 (X12739) and CAMR5420 (X5420). Successful intracellular expression of the four recombinant proteins in *E. coli* was achieved and the enzymes were shown to be active in both, minimal media containing xylan plates, and reducing-sugars DNS assays, with XC56 showing the highest xylanolytic activity followed by X12739 and X5420. As expected, IXC56 showed very little activity against xylan, but was most active with *p*NP-β-D-xylopyranoside as substrate, confirming that this intracellular enzyme behaves more as a β-xylosidase than a xylanase.

The extracellular xylanase-encoding genes were also cloned in an expression vector for heterologous constitutive expression in *Geobacillus*. Successful expression in *G. thermoglucosidasius* TM242, and secretion of the heterologous extracellular xylanase

XC56 in an active form, were achieved. It is unclear why the expression and secretion of the other two extracellular xylanases was unsuccessful, and with more time the experiments would have been repeated to achieve the cloning and expression of X12739 and X5420 in TM242. However, as XC56 appeared to be catalytically superior to the other two xylanases when expressed in *E. coli,* it can be suggested that even successful expression and secretion of X12739 and X5420 in *G. thermoglucosidasius* TM242 will not be better than the results obtained with XC56.

Although quantitative analysis of the heterologous XC56 xylanolytic activity was not possible in the timescale of this project, further characterisation of the XC56 enzyme expressed in both *E. coli* and TM242 is currently under investigation by the PhD candidate Alexandria Holland (University of Bath), who is working on the purification of the enzyme, assaying its xylanolytic activity in a quantitative manner (in order to obtain the enzyme kinetics) and planning to study the protein secretion in *Geobacillus*.

6.3. Future work

A study of the synergy between all the different xylan-degrading enzymes obtained during this project, and of their evaluation in large-scale ethanol-producing fermentation facilities (monitored to determine the biochemical basis of any industrially significant changes observed), is needed in order to assess the contribution of the findings presented here towards the improvement of second-generation bioethanol production

It can be suggested that the TM242 intracellular enzymes, in particular β-xylosidase-1 and α-arabinofuranoside (both with higher catalytic activity than the enzymes from *G. stearothermophilus* T-6), already work in a very efficient manner inside *G. thermoglucosidasius* TM242. However, enhancement of the enzymatic hydrolysis of hemicellulosic feedstock prior to fermentation might be achieved by structurally-based engineering of these enzymes and/or their upregulation in the TM242 genome. As demonstrated by Cripps *et al.* (2009), successful upregulation can be achieved by replacing the native promoter region upstream of the target genes with anaerobically inducible promoters (*e.g. ldh* promoters from *G. stearothermophilus* NCA1503 or the *pfl* promoter from *Bacillus cereus* ATCC14579).

In addition, process improvement would be also gained by enhanced secretion of the intracellular xylan-degrading enzymes from TM242. It has been shown that successful

extracellular expression of intracellular glycoside hydrolases can be achieved by cloning the target gene downstream of a known signal peptide sequence (predicted with high confidence as a Sec system signal peptide) into a *Geobacillus* expression vector (Bartosiak-Jentys *et al.*, 2013).

On the other hand, the integration (knock-in) of the missing xylanase-encoding gene(s) into the TM242 genome might effect enhanced biomass utilisation by this microorganism, contributing towards the development of consolidated bioprocessing of lignocellulose to bioethanol. If successful, this will improve the technology currently under development by ReBio Technologies Ltd and will add considerably to their ability to commercialise their bioethanol process.

The knock-in of heterologous genes has been achieved before,where, as detailed in a US Patent filled by TMO Renewables Ltd (2012), an α-amylase-encoding gene from *G. stearothermophilus* DSM22 was successfully integrated into the previously knocked-out TM242 *pfl* locus (to form strain TM333). As suggested by Dr Gareth Cooper (personal communication), the knock-in of heterologous xylanases into TM242 can be done using a modification of the basic gene exchange approach using double recombination. In this approach, a non-essential or unwanted gene in the genome of the microorganism is used for the insertion of the new heterologous gene. First, the heterologous gene needs to be cloned with a suitable promoter upstream into a vector such as pUC19; then, some DNA sequence from TM242 (from the locus where the gene is going to be inserted) is cloned either side of the heterologous gene. This cassette is then cloned into a knock-in plasmid, becoming the completed knock-in construct which is then transformed into TM242. Finally, primary integrants (single cross-overs) need to be selected, from which the finished knock-in (double cross-overs) containing the heterologous gene but not the other DNA from the plasmid, can be obtained.

To conclude, the work presented in this thesis seeks to further the understanding of enzyme hydrolysis of renewable biomass, in particular of hemicellulose, and to aid the optimisation of this step during second-generation bioethanol production. Hopefully the findings reported here will be studied further and exploited at ReBio Technologies Ltd.

REFERENCES

Adams, P.D., Afonine, P. V., Bunkóczi, G., *et al.* (2010). PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. Sect. D, Biol. Crystallogr.* **66**, 213–221.

Aleshin, A.E., Feng, P.-H., Honzatko, R.B., & Reilly, P.J. (2003). Crystal structure and evolution of a prokaryotic glucoamylase. *J. Mol. Biol.* **327**, 61–73.

Anand, A., Kumar, V., & Satyanarayana, T. (2013). Characteristics of thermostable endoxylanase and β-xylosidase of the extremely thermophilic bacterium *Geobacillus thermodenitrificans* TSAA1 and its applicability in generating xylooligosaccharides and xylose from agro-residues. *Extrem. Life under Extrem. Cond.* **17**, 357–366.

Arantes, V., & Saddler, J.N. (2010). Access to cellulose limits the efficiency of enzymatic hydrolysis: the role of amorphogenesis. *Biotechnol. Biofuels* **3**, 4.

Atkins, G.L., & Nimmo, I.A. (1975). A comparison of seven methods for fitting the Michaelis-Menten equation. *Biochem. J.* **149**, 775–777.

Bailey, M.J., Biely, P., & Poutanen, K. (1992). Interlaboratory testing of methods for assay of xylanase activity. *J. Biotechnol.* **23**, 257–270.

Bartosiak-Jentys, J., Hussein, A.H., Lewis, C.J., & Leak, D.J. (2013). Modular system for assessment of glycosyl hydrolase secretion in *Geobacillus thermoglucosidasius*. *Microbiology* **159**, 1267–1275.

Berman, H.M. (2000). The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242.

Blow, D. (2002). Outline of Crystallography for Biologists. Oxford University Press Inc., Oxford, UK.

Blumer-Schuette, S.E., Brown, S.D., Sander, K.B., Bayer, E.A., Kataeva, I., Zurawski, J. V, Conway, J.M., Adams, M.W.W., & Kelly, R.M. (2014). Thermophilic lignocellulose deconstruction. *FEMS Microbiol. Rev.* **38**, 393–448.

Boggon, T.J., & Shapiro, L. (2000). Screening for phasing atoms in protein crystallography. *Structure* **8**, R143–R149.

Bokinsky, G., Peralta-Yahya, P.P., George, A., Holmes, B.M., Steen, E.J., Dietrich, J., Lee, T.S., Tullman-Ercek, D., Voigt, C.A., Simmons, B.A., & Keasling, J.D. (2011). Synthesis of three advanced biofuels from ionic liquid-pretreated switchgrass using engineered *Escherichia coli*. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 19949–19954.

BP (2014a). Statistical Review of World Energy June 2014 (63th edn). BP p.l.c., London, UK.

BP (2014b). Energy Outlook 2035 January 2014. BP p.l.c., London, UK.

Bradford, M.M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* **72**, 248–254.

Bravman, T., Zolotnitsky, G., Shulami, S., Belakhov, V., Solomon, D., Baasov, T., Shoham, G., & Shoham, Y. (2001). Stereochemistry of family 52 glycosyl hydrolases: A β-xylosidase from *Bacillus stearothermophilus* T-6 is a retaining enzyme. *FEBS Lett.* **495**, 39–43.

Bravman, T., Zolotnitsky, G., Belakhov, V., Shoham, G., Henrissat, B., Baasov, T., & Shoham, Y. (2003). Detailed kinetic analysis of a family 52 glycoside hydrolase: A β-xylosidase from *Geobacillus stearothermophilus*. *Biochemistry* **42**, 10528–10536.

Brunzelle, J.S., Jordan, D.B., McCaslin, D.R., Olczak, A., & Wawrzak, Z. (2008). Structure of the two-subsite β-D-xylosidase from *Selenomonas ruminantium* in complex with 1,3-bis[tris(hydroxymethyl)methylamino]propane. *Arch. Biochem. Biophys.* **474**, 157–166.

Brüx, C., Ben-David, A., Shallom-Shezifi, D., Leon, M., Niefind, K., Shoham, G., Shoham, Y., & Schomburg, D. (2006). The structure of an inverting GH43 β-xylosidase from *Geobacillus stearothermophilus* with its substrate reveals the role of the three catalytic residues. *J. Mol. Biol.* **359**, 97–109.

Canakci, S., Belduz, A.O., Saha, B.C., Yasar, A., Ayaz, F.A., & Yayli, N. (2007). Purification and characterization of a highly thermostable α-L-Arabinofuranosidase from Geobacillus caldoxylolyticus TK4. *Appl. Microbiol. Biotechnol.* **75**, 813–820.

Canakci, S., Cevher, Z., Inan, K., Tokgoz, M., Bahar, F., Kacagan, M., Sal, F.A., & Belduz, A.O. (2012). Cloning, purification and characterization of an alkali-stable endoxylanase from thermophilic Geobacillus sp. 71. *World J. Microbiol. Biotechnol.* **28**, 1981–1988.

Chen, V.B., Arendall, W.B., Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., & Richardson, D.C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. Sect. D, Biol. Crystallogr.* **66**, 12–21.

Cherubini, F., & Strømman, A.H. (2011). Life cycle assessment of bioenergy systems: state of the art and future challenges. *Bioresour. Technol.* **102**, 437–451.

Chiaramonti, D., Prussi, M., Ferrero, S., Oriani, L., Ottonello, P., Torre, P., & Cherchi, F. (2012). Review of pretreatment processes for lignocellulosic ethanol production, and development of an innovative method. *Biomass and Bioenergy* **46**, 25–35.

Christopher, L.P., Yao, B., & Ji, Y. (2014). Lignin Biodegradation with Laccase-Mediator Systems. *Front. Energy Res.* **2**.

Chung, D., Cha, M., Guss, A.M., & Westpheling, J. (2014). Direct conversion of plant biomass to ethanol by engineered *Caldicellulosiruptor bescii*. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 8931–8936.

Collins, T., Gerday, C., & Feller, G. (2005). Xylanases, xylanase families and extremophilic xylanases. *FEMS Microbiol. Rev.* **29**, 3–23.

Contreras, L.M., Gómez, J., Prieto, J., Clemente-Jiménez, J.M., Las Heras-Vázquez, F.J., Rodríguez-Vico, F., Blanco, F.J., & Neira, J.L. (2008). The family 52 β-xylosidase from *Geobacillus stearothermophilus* is a dimer: structural and biophysical characterization of a glycoside hydrolase. *Biochim. Biophys. Acta - Proteins Proteomics* **1784**, 1924–1934.

Cripps, R.E., Eley, K., Leak, D.J., Rudd, B., Taylor, M., Todd, M., Boakes, S., Martin, S., & Atkinson, T. (2009). Metabolic engineering of *Geobacillus thermoglucosidasius* for high yield ethanol production. *Metab. Eng.* **11**, 398–408.

Cronin, V.A. (2013). Recombinant Expression, Purification and Characterization of a glycoside hydrolase family 43 β-Xylosidase and a family 52 β-Xylosidase from *Geobacillus themoglucosidasius*. Project Report for the Degree of BSc (Hons) Biochemistry. University of Bath, Bath, UK.

Czjzek, M., Bravman, T., Henrissat, B., & Shoham, Y. (2004). Crystallization and preliminary crystallographic analysis of a thermostable family 52 β-D-xylosidase from *Geobacillus stearothermophilus* T-6. *Acta Crystallogr. Sect. D, Biol. Crystallogr.* **60**, 1461–1463.

Czjzek, M., Ben David, A., Bravman, T., Shoham, G., Henrissat, B., & Shoham, Y. (2005). Enzyme-substrate complex structures of a GH39 β-xylosidase from *Geobacillus stearothermophilus*. *J. Mol. Biol.* **353**, 838–846.

Davies, G., & Henrissat, B. (1995). Structures and mechanisms of glycosyl hydrolases. *Structure* **3**, 853–859.

Debeche, T., Cummings, N., Connerton, I., Debeire, P., & O'Donohue, M.J. (2007). Genetic and biochemical characterization of a highly thermostable α-L-arabinofuranosidase from *Thermobacillus xylanilyticus*. *Appl. Environ. Microbiol.* **66**, 1734–1736.

Demain, A.L., Newcomb, M., & Wu, J.H.D. (2005). Cellulase, clostridia, and ethanol. *Microbiol. Mol. Biol. Rev.* **69**, 124–154.

Demirbas, A. (2009). Political, economic and environmental impacts of biofuels: A review. *Appl. Energy* **86**, S108–S117.

Department for Transport (2007). Renewable Transport Fuels Obligation (RTFO) Order 2007 (S.I. 2007/3072), London, UK.

Dodson, E. (2003). Is it jolly SAD? *Acta Crystallogr. Sect. D, Biol. Crystallogr.* **56**, 1958–1965.

Donovan, R.S., Robinson, C.W., & Glick, B.R. (1996). Review: optimizing inducer and culture conditions for expression of foreign proteins under the control of the *lac* promoter. *J. Ind. Microbiol.* **16**, 145–154.

Doublié, S. (2007). Production of selenomethionyl proteins in prokaryotic and eukaryotic expression systems. *Methods Mol. Biol.* **363**, 91–108.

Drenth, J. (1994). Principles of Protein X-ray Crystallography. Springer-Verlag, New York, US.

Eisenthal, R., & Cornish-Bowden, A. (1974). The direct linear plot. A new graphical procedure for estimating enzyme kinetic parameters. *Biochem. J.* **139**, 715–720.

Emsley, P., Lohkamp, B., Scott, W.G., & Cowtan, K. (2010). Features and development of Coot. *Acta Crystallogr. Sect. D, Biol. Crystallogr.* **66**, 486–501.

Eram, M.S., & Ma, K. (2013). Decarboxylation of pyruvate to acetaldehyde for ethanol production by hyperthermophiles. *Biomolecules* **3**, 578–596.

Espina, G., Eley, K., Pompidor, G., Schneider, T.R., Crennell, S.J., & Danson, M.J. (2014). A novel β-xylosidase structure from *Geobacillus thermoglucosidasius*: the first crystal structure of a glycoside hydrolase family GH52 enzyme reveals unpredicted similarity to other glycoside hydrolase folds. *Acta Crystallogr. Sect. D, Biol. Crystallogr.* **70**, 1366–1374.

Evans, P. (2006). Scaling and assessment of data quality. *Acta Crystallogr. Sect. D, Biol. Crystallogr.* **62**, 72–82.

Gaberc-Porekar, V., & Menart, V. (2001). Perspectives of immobilized-metal affinity chromatography. *J. Biochem. Biophys. Methods* **49**, 335–360.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D., & Bairoch, A. (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* **31**, 3784–3788.

Ghaffar, S.H., & Fan, M. (2013). Structural analysis for lignin characteristics in biomass straw. *Biomass and Bioenergy* **57**, 264–279.

Gilead, S., & Shoham, Y. (1995). Purification and characterization of α-L-arabinofuranosidase from *Bacillus stearothermophilus* T-6. *Appl. Environ. Microbiol.* **61**, 170–174.

Goffin, E. (2014). Purification and Characterisation of a Glycoside Hydrolase Family 51 α-L-Arabinofuranosidase from Geobacillus thermoglucosidasius. Project Report for the Degree of BSc (Hons) Biochemistry. University of Bath, Bath, UK.

Goodsell, D.S. (2010). The Protein Data Bank: Exploring Biomolecular Structure. Nature Education. *Nat. Educ.* **3**, 39.

Gouet, P., Courcelle, E., Stuart, D.I., & Métoz, F. (1999). ESPript: analysis of multiple sequence alignments in PostScript. *Bioinformatics* **15**, 305–308.

Grabski, A., Mehler, M., & Drott, D. (2005). The Overnight Express Autoinduction System: High-density cell growth and protein expression while you sleep. *Nat. Methods* **2**, 233–235.

Grodberg, J., & Dunn, J.J. (1988). ompT encodes the *Escherichia coli* outer membrane protease that cleaves T7 RNA polymerase during purification. *J. Bacteriol.* **170**, 1245–1253.

Gusakov, A. V, Kondratyeva, E.G., & Sinitsyn, A.P. (2011). Comparison of two methods for assaying reducing sugars in the determination of carbohydrase activities. *Int. J. Anal. Chem.* **2011**, 283658.

Haghighi Mood, S., Hossein Golfeshan, A., Tabatabaei, M., Salehi Jouzani, G., Najafi, G.H., Gholami, M., & Ardjmand, M. (2013). Lignocellulosic biomass to bioethanol, a comprehensive review with a focus on pretreatment. *Renew. Sustain. Energy Rev.* **27**, 77–93.

Hasunuma, T., Okazaki, F., Okai, N., Hara, K.Y., Ishii, J., & Kondo, A. (2013). A review of enzymes and microbes for lignocellulosic biorefinery and the possibility of their application to consolidated bioprocessing technology. *Bioresour. Technol.* **135**, 513–522.

Henrissat, B., Callebaut, I., Fabrega, S., Lehn, P., Mornon, J.P., & Davies, G. (1995). Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 7090–7094.

Himmel, M.E., Ding, S.-Y., Johnson, D.K., Adney, W.S., Nimlos, M.R., Brady, J.W., & Foust, T.D. (2007). Biomass recalcitrance: engineering plants and enzymes for biofuels production. *Science* **315**, 804–807.

Holm, L., & Rosenström, P. (2010). Dali server: conservation mapping in 3D. *Nucleic Acids Res.* **38**, W545–W549.

Hövel, K., Shallom, D., Niefind, K., Baasov, T., Shoham, G., Shoham, Y., & Schomburg, D. (2003a). Crystallization and preliminary X-ray analysis of a family 51 glycoside hydrolase, the α-L-arabinofuranosidase from *Geobacillus stearothermophilus* T-6. *Acta Crystallogr. Sect. D, Biol. Crystallogr.* **59**, 913–915.

Hövel, K., Shallom, D., Niefind, K., Belakhov, V., Shoham, G., Baasov, T., Shoham, Y., & Schomburg, D. (2003b). Crystal structure and snapshots along the reaction pathway of a family 51 α-L-arabinofuranosidase. *EMBO J.* **22**, 4922–4932.

Huang, C.H., Sun, Y., Ko, T.P., Chen, C.C., Zheng, Y., Chan, H.C., Pang, X., Wiegel, J., Shao, W., & Guo, R.T. (2012). The substrate/product-binding modes of a novel GH120 β-xylosidase (XylC) from *Thermoanaerobacterium saccharolyticum* JW/SL-YS485. *Biochem. J.* **448**, 401–407.

Huang, Z., Liu, X., Zhang, S., & Liu, Z. (2014). GH52 xylosidase from *Geobacillus stearothermophilus*: characterization and introduction of xylanase activity by site-directed mutagenesis of Tyr509. *J. Ind. Microbiol. Biotechnol.* **41**, 65–74.

IPCC (2014). Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. IPCC, Geneva, Switzerland.

Jeffries, T.W., Yang, V.W., & Davis, M.W. (1998). Comparative study of xylanase kinetics using dinitrosalicylic, arsenomolybdate, and ion chromatographic assays. *Appl. Biochem. Biotechnol.* **70-72**, 257–265.

Jeng, W.-Y., Wang, N.-C., Lin, M.-H., Lin, C.-T., Liaw, Y.-C., Chang, W.-J., Liu, C.-I., Liang, P.-H., & Wang, A.H.-J. (2011). Structural and functional analysis of three β-glucosidases from bacterium *Clostridium cellulovorans*, fungus *Trichoderma reesei* and termite Neotermes koshunensis. *J. Struct. Biol.* **173**, 46–56.

Jenkins, J., Lo Leggio, L., Harris, G., & Pickersgill, R. (1995). β-glucosidase, β-galactosidase, family A cellulases, family F xylanases and two barley glycanases form a superfamily of enzymes with 8-fold β/α architecture and with two conserved glutamates near the carboxy-terminal ends of β-strands four. *FEBS Lett.* **362**, 281–285.

Jordan, D.B., & Li, X.-L. (2007). Variation in relative substrate specificity of bifunctional β-D-xylosidase/α-L-arabinofuranosidase by single-site mutations: roles of substrate distortion and recognition. *Biochim. Biophys. Acta* **1774**, 1192–1198.

Jordan, D.B., & Wagschal, K. (2010). Properties and applications of microbial β-D-xylosidases featuring the catalytically efficient enzyme from *Selenomonas ruminantium*. *Appl. Microbiol. Biotechnol.* **86**, 1647–1658.

Kabsch, W. (2010). XDS. *Acta Crystallogr. Sect. D, Biol. Crystallogr.* **66**, 125–132.

Khasin, A., Alchanati, I., & Shoham, Y. (1993). Purification and characterization of a thermostable xylanase from *Bacillus stearothermophilus* T-6. *Appl. Environ. Microbiol.* **59**, 1725–1730.

Klein-Marcuschamer, D., Oleskowicz-Popiel, P., Simmons, B.A., & Blanch, H.W. (2012). The challenge of enzyme cost in the production of lignocellulosic biofuels. *Biotechnol. Bioeng.* **109**, 1083–1087.

Koshland, D.E. (1953). Stereochemistry and the mechanism of enzymic reactions. *Biol. Revs. Cambridge Phil. Soc.* **28**, 416–436.

Krissinel, E., & Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. Sect. D, Biol. Crystallogr.* **60**, 2256–2268.

Krissinel, E., & Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797.

Kumar, D., & Murthy, G.S. (2011). Impact of pretreatment and downstream processing technologies on economics and energy in cellulosic ethanol production. *Biotechnol. Biofuels* **4**, 27.

Lal, R. (2008). Crop residues as soil amendments and feedstock for bioethanol production. *Waste Manag.* **28**, 747–758.

Langer, G., Cohen, S.X., Lamzin, V.S., & Perrakis, A. (2008). Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat. Protoc.* **3**, 1171–1179.

Lee, R.A., & Lavoie, J.-M. (2013). From first- to third-generation biofuels: Challenges of producing a commodity from a biomass of increasing complexity. *Anim. Front.* **3**, 6–11.

Lee, T.H., Lim, P.O., & Lee, Y.E. (2007). Purification and Characterization of β-Xylosidase from Paenibacillus sp. DG-22. *J. Life Sci.* **17**, 1341–1346.

Litzinger, S., Fischer, S., Polzer, P., Diederichs, K., Welte, W., & Mayer, C. (2010). Structural and kinetic analysis of *Bacillus subtilis* N-acetylglucosaminidase reveals a unique Asp-His dyad mechanism. *J. Biol. Chem.* **285**, 35675–35684.

Liu, B., Zhang, N., Zhao, C., Lin, B., Xie, L., & Huang, Y. (2012). Characterization of a recombinant thermostable xylanase from hot spring thermophilic *Geobacillus* sp. TC-W7. *J. Microbiol. Biotechnol.* **22**, 1388–1394.

Liu, L., Cheng, S.Y., Li, J.B., & Huang, Y.F. (2007). Mitigating Environmental Pollution and Impacts from Fossil Fuels: The Role of Alternative Fuels. *Energy Sources, Part A Recover. Util. Environ. Eff.* **29**, 1069–1080.

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P.M., & Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490–D495.

Lovell, S.C., Davis, I.W., Arendall, W.B., de Bakker, P.I.W., Word, J.M., Prisant, M.G., Richardson, J.S., & Richardson, D.C. (2003). Structure validation by Cα geometry: ϕ,ψ and Cβ deviation. *Proteins* **50**, 437–450.

Lynd, L.R., Weimer, P.J., van Zyl, W.H., & Pretorius, I.S. (2002). Microbial cellulose utilization: fundamentals and biotechnology. *Microbiol. Mol. Biol. Rev.* **66**, 506–577, table of contents.

Lynd, L.R., van Zyl, W.H., McBride, J.E., & Laser, M. (2005). Consolidated bioprocessing of cellulosic biomass: an update. *Curr. Opin. Biotechnol.* **16**, 577–583.

De Maayer, P., Brumm, P.J., Mead, D.A., & Cowan, D.A. (2014a). Comparative analysis of the *Geobacillus* hemicellulose utilization locus reveals a highly variable target for improved hemicellulolysis. *BMC Genomics* **15**, 836.

De Maayer, P., Williamson, C.E., Vennard, C.T., Danson, M.J., & Cowan, D.A. (2014b). Draft Genome Sequences of *Geobacillus* sp. Strains CAMR5420 and CAMR12739. *Genome Announc.* **2**, e00567–14.

Marchler-Bauer, A., Zheng, C., Chitsaz, F., *et al.* (2013). CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res.* **41**, D348–D352.

McCarter, J.D., & Withers, S.G. (1994). Mechanisms of enzymatic glycoside hydrolysis. *Curr. Opin. Struct. Biol.* **4**, 885–892.

McKee, L.S., Peña, M.J., Rogowski, A., Jackson, A., Lewis, R.J., York, W.S., Krogh, K.B.R.M., Viksø-Nielsen, A., Skjøt, M., Gilbert, H.J., & Marles-Wright, J. (2012). Introducing endo-xylanase activity into an exo-acting arabinofuranosidase that targets side chains. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 6537–6542.

Menon, V., & Rao, M. (2012). Trends in bioconversion of lignocellulose: Biofuels, platform chemicals & biorefinery concept. *Prog. Energy Combust. Sci.* **38**, 522–550.

Miller, G.L. (1959). Use of Dinitrosalicylic Acid Reagent for Determination of Reducing Sugar. *Anal. Chem.* **31**, 426–428.

Minic, Z., Rihouey, C., Do, C.T., Lerouge, P., & Jouanin, L. (2004). Purification and characterization of enzymes exhibiting β-D-xylosidase activities in stem tissues of *Arabidopsis*. *Plant Physiol.* **135**, 867–878.

Morales, M., Quintero, J., Conejeros, R., & Aroca, G. (2015). Life cycle assessment of lignocellulosic bioethanol: Environmental impacts and energy balance. *Renew. Sustain. Energy Rev.* **42**, 1349–1361.

Murshudov, G.N., Skubák, P., Lebedev, A.A., Pannu, N.S., Steiner, R.A., Nicholls, R.A., Winn, M.D., Long, F., & Vagin, A.A. (2011). REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. Sect. D, Biol. Crystallogr.* **67**, 355–367.

Mussatto, S.I., Dragone, G., Guimarães, P.M.R., Silva, J.P.A., Carneiro, L.M., Roberto, I.C., Vicente, A., Domingues, L., & Teixeira, J.A. (2010). Technological trends, global market, and challenges of bio-ethanol production. *Biotechnol. Adv.* **28**, 817–830.

Nanmori, T., Watanabe, T., Shinke, R., Kohno, A., & Kawamura, Y. (1990). Purification and properties of thermostable xylanase and β-xylosidase produced by a newly isolated *Bacillus stearothermophilus* strain. *J. Bacteriol.* **172**, 6669–6672.

Novagen (2005). pET System Manual 11th edition. EMD Biosciences Inc., Darmstadt, Germany.

NREL (2007). A National Laboratory Market and Technology Assessment of the 30x30 Scenario Technical DOE Report, NREL TP-510-40942. National Renewable Energy Laboratory, Colorado, US.

OECD/FAO (2014). OECD-FAO Agricultural Outlook 2014. OECD Publishing, Paris, France.

Olson, D.G., McBride, J.E., Shaw, A.J., & Lynd, L.R. (2012). Recent progress in consolidated bioprocessing. *Curr. Opin. Biotechnol.* **23**, 396–405.

Otwinowski, Z., & Minor, W. (1997). Macromolecular Crystallography Part A. *Methods Enzymol.* **276**, 307–326.

Overbeek, R., Larsen, N., Walunas, T., *et al.* (2003). The ERGO genome analysis and discovery system. *Nucleic Acids Res.* **31**, 164–171.

Pape, T., & Schneider, T.R. (2004). HKL2MAP: a graphical user interface for macromolecular phasing with SHELX programs. *J. Appl. Crystallogr.* **37**, 843–844.

Petersen, T.N., Brunak, S., von Heijne, G., & Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* **8**, 785–786.

Petrova, P., & Ivanova, V. (2014). Perspectives for the Production of Bioethanol from Lignocellulosic Materials. *Biotechnol. Biotechnol. Equip.* **24**, 529–546.

Pollet, A., Delcour, J.A., & Courtin, C.M. (2010). Structural determinants of the substrate specificities of xylanases from different glycoside hydrolase families. *Crit. Rev. Biotechnol.* **30**, 176–191.

Pompidor, G., D'Aléo, A., Vicat, J., Toupet, L., Giraud, N., Kahn, R., & Maury, O. (2008). Protein crystallography through supramolecular interactions between a lanthanide complex and arginine. *Angew. Chem. Int. Ed. Engl.* **47**, 3388–3391.

Quintero, D., Velasco, Z., Hurtado-Gómez, E., Neira, J.L., & Contreras, L.M. (2007). Isolation and characterization of a thermostable β-xylosidase in the thermophilic bacterium *Geobacillus pallidus*. *Biochim. Biophys. Acta* **1774**, 510–518.

Rabinovitch-Deere, C.A., Oliver, J.W.K., Rodriguez, G.M., & Atsumi, S. (2013). Synthetic biology and metabolic engineering approaches to produce biofuels. *Chem. Rev.* **113**, 4611–4632.

Ratanakhanokchai, K., Waeonukul, R., Pason, P., Tachaapaikoon, C., Khin, L.K., Sakka, K., Kosugi, A., & Mori, Y. (2013). In: Biomass Now - Cultivation and Utilization, Matovic, M. D., Ed.; InTech, Rijeka, Croatia.

Rhodes, G. (1993). Crystallography Made Crystal Clear: A Guide for Users of Macromolecular Models. Academic Press Limited, London, UK.

Rice, L.M., Earnest, T.N., & Brunger, A.T. (2000). Single-wavelength anomalous diffraction phasing revisited. *Acta Crystallogr. Sect. D, Biol. Crystallogr.* **56**, 1413–1420.

Royer, S.F., Haslett, L., Crennell, S.J., Hough, D.W., Danson, M.J.& Bull, S.D. (2010). Structurally informed site-directed mutagenesis of a stereochemically promiscuous aldolase to afford stereochemically complementary biocatalysts. *J. Am. Chem. Soc.* **132**, 11753–11758.

Saha, B.C. (2003). Hemicellulose bioconversion. *J. Ind. Microbiol. Biotechnol.* **30**, 279–291.

Saha, B.C., Jordan, D.B., & Bothast, R.J. (2009). Enzymes, Industrial (Overview). In: Schaechter, M., Ed; Encyclopedia of Microbiology. Elsevier, Oxford, UK.

Sambrook, J., & Russell, D.W. (2001). Molecular Cloning: A laboratory manual - Vol. 1, 2, 3, 3rd edn. Cold Spring Harbor Laboratory Press, New York, US.

Sánchez, O.J., & Cardona, C.A. (2008). Trends in biotechnological production of fuel ethanol from different feedstocks. *Bioresour. Technol.* **99**, 5270–5295.

Santos, C.R., Polo, C.C., Corrêa, J.M., Simão, R. de C.G., Seixas, F.A.V., & Murakami, M.T. (2012). The accessory domain changes the accessibility and molecular topography of the catalytic interface in monomeric GH39 β-xylosidases. *Acta Crystallogr. Sect. D, Biol. Crystallogr.* **68**, 1339–1345.

Scheller, H.V., & Ulvskov, P. (2010). Hemicelluloses. *Annu. Rev. Plant Biol.* **61**, 263–289.

Shallom, D., Belakhov, V., Solomon, D., Shoham, G., Baasov, T., & Shoham, Y. (2002a). Detailed kinetic analysis and identification of the nucleophile in α-L-arabinofuranosidase from *Geobacillus stearothermophilus* T-6, a family 51 glycoside hydrolase. *J. Biol. Chem.* **277**, 43667–43673.

Shallom, D., Belakhov, V., Solomon, D., Gilead-Gropper, S., Baasov, T., Shoham, G., & Shoham, Y. (2002b). The identification of the acid–base catalyst of α-arabinofuranosidase from *Geobacillus stearothermophilus* T-6, a family 51 glycoside hydrolase. *FEBS Lett.* **514**, 163–167.

Sheehan, J.J. (2009). Biofuels and the conundrum of sustainability. *Curr. Opin. Biotechnol.* **20**, 318–324.

Sheldrick, G.M. (2008). A short history of SHELX. *Acta Crystallogr. Sect. A, Found. Crystallogr.* **64**, 112–122.

Shulami, S., Zaide, G., Zolotnitsky, G., Langut, Y., Feld, G., Sonenshein, A.L., & Shoham, Y. (2007). A two-component system regulates the expression of an ABC transporter for xylo-oligosaccharides in *Geobacillus stearothermophilus*. *J. Bacteriol.* **73**, 874–884.

Shulami, S., Raz-Pasteur, A., Tabachnikov, O., Gilead-Gropper, S., Shner, I., & Shoham, Y. (2011). The L-Arabinan utilization system of *Geobacillus stearothermophilus*. *J. Bacteriol.* **193**, 2838–2850.

Shulami, S., Shenker, O., Langut, Y., Lavid, N., Gat, O., Zaide, G., Zehavi, A., Sonenshein, A.L., & Shoham, Y. (2014). Multiple regulatory mechanisms control the expression of the *Geobacillus stearothermophilus* gene for extracellular xylanase. *J. Biol. Chem.* **289**, 25957–25975.

Simmons, B.A., Loque, D., & Blanch, H.W. (2008). Next-generation biomass feedstocks for biofuel production. *Genome Biol.* **9**, 242.

Solomon, V., Teplitsky, A., Shulami, S., Zolotnitsky, G., Shoham, Y., & Shoham, G. (2007). Structure-specificity relationships of an intracellular xylanase from *Geobacillus stearothermophilus*. *Acta Crystallogr. Sect. D, Biol. Crystallogr.* **63**, 845–859.

Stepper, J., Dabin, J., Eklof, J.M., Thongpoo, P., Kongsaeree, P., Taylor, E.J., Turkenburg, J.P., Brumer, H., & Davies, G.J. (2013). Structure and activity of the *Streptococcus pyogenes* family GH1 6-phospho-β-glucosidase SPy1599. *Acta Crystallogr. Sect. D, Biol. Crystallogr.* **69**, 16–23.

Studier, F.W. (2005). Protein production by auto-induction in high-density shaking cultures. *Protein Expr. Purif.* **41**, 207–234.

Studier, F.W., & Moffatt, B.A. (1986). Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *J. Mol. Biol.* **189**, 113–130.

Sun, R. (2010). Cereal Straw as a Resource for Sustainable Biomaterials and Biofuels: Chemistry, Extractives, Lignins, Hemicelluloses and Cellulose. Elsevier, Boston, US.

Sun, Y., & Cheng, J. (2002). Hydrolysis of lignocellulosic materials for ethanol production: a review. *Bioresour. Technol.* **83**, 1–11.

Suzuki, H., Yoshida, K., & Ohshima, T. (2013). Polysaccharide-degrading thermophiles generated by heterologous gene expression in *Geobacillus kaustophilus* HTA426. *Appl. Environ. Microbiol.* **79**, 5151–5158.

Suzuki, T., Kitagawa, E., Sakakibara, F., Ibata, K., Usui, K., & Kawai, K. (2001). Cloning, expression, and characterization of a family 52 β-xylosidase gene (xysB) of a multiple-xylanase-producing bacterium, *Aeromonas caviae* ME-1. *Biosci. Biotechnol. Biochem.* **65**, 487–494.

Taylor, G.L. (2003). The phase problem. *Acta Crystallogr. Sect. D, Biol. Crystallogr.* **59**, 1881–1890.

Taylor, G.L. (2010). Introduction to phasing. *Acta Crystallogr. Sect. D, Biol. Crystallogr.* **66**, 325–338.

Taylor, M.P., Esteban, C.D., & Leak, D.J. (2008). Development of a versatile shuttle vector for gene expression in *Geobacillus* spp. *Plasmid* **60**, 45–52.

Teather, R.M., & Wood, P.J. (1982). Use of Congo red-polysaccharide interactions in enumeration and characterization of cellulolytic bacteria from the bovine rumen. *Appl. Environ. Microbiol.* **43**, 777–780.

Teplitsky, A., Mechaly, A., Stojanoff, V., Sainz, G., Golan, G., Feinberg, H., Gilboa, R., Reiland, V., Zolotnitsky, G., Shallom, D., Thompson, A., Shoham, Y., & Shoham, G. (2004). Structure determination of the extracellular xylanase from Geobacillus stearothermophilus by selenomethionyl MAD phasing. *Acta Crystallogr. Sect. D, Biol. Crystallogr.* **60**, 836–848.

The European Parliament and the Council of the European Union (2009). Directive 2009/28/EC of the European Parliament and of the Council of 23 April 2009 on the promotion of the use of energy from renewable sources and amending and subsequently repealing Directives 2001/77/EC and 2003/30/EC. *Off. J. Eur. Union*, **L140**, 16–62.

TMO Renewables Ltd, Atkinson, A., Cripps, R., Eley, K., Rudd, B., & Todd, M. (2012). Thermophilic microorganisms for ethanol production. US Patent: US8143038 B2.

Vanholme, R., Van Acker, R., & Boerjan, W. (2010). Potential of Arabidopsis systems biology to advance the biofuel field. *Trends Biotechnol.* **28**, 543–547.

Viikari, L., Vehmaanperä, J., & Koivula, A. (2012). Lignocellulosic ethanol: From science to industry. *Biomass and Bioenergy* **46**, 13–24.

Vuong, T. V, & Wilson, D.B. (2010). Glycoside hydrolases: catalytic base/nucleophile diversity. *Biotechnol. Bioeng.* **107**, 195–205.

Wagschal, K., Heng, C., Lee, C.C., & Wong, D.W.S. (2009). Biochemical characterization of a novel dual-function arabinofuranosidase/xylosidase isolated from a compost starter mixture. *Appl. Microbiol. Biotechnol.* **81**, 855–863.

Wang, Y., Feng, S., Zhan, T., Huang, Z., Wu, G., & Liu, Z. (2013). Improving catalytic efficiency of endo-β-1, 4-xylanase from *Geobacillus stearothermophilus* by directed evolution and H179 saturation mutagenesis. *J. Biotechnol.* **168**, 341–347.

Weickert, M.J., Doherty, D.H., Best, E.A., & Olins, P.O. (1996). Optimization of heterologous protein production in *Escherichia coli*. *Curr. Opin. Biotechnol.* **7**, 494–499.

Wlodawer, A., Minor, W., Dauter, Z., & Jaskolski, M. (2008). Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *FEBS J.* **275**, 1–21.

Yang, J.K., Yoon, H.J., Ahn, H.J., Lee, B. Il, Pedelacq, J.-D., Liong, E.C., Berendzen, J., Laivenieks, M., Vieille, C., Zeikus, G.J., Vocadlo, D.J., Withers, S.G., & Suh, S.W. (2004). Crystal structure of β-D-xylosidase from *Thermoanaerobacterium saccharolyticum*, a family 39 glycoside hydrolase. *J. Mol. Biol.* **335**, 155–165.

Yang, X., Shi, P., Huang, H., Luo, H., Wang, Y., Zhang, W., & Yao, B. (2014). Two xylose-tolerant GH43 bifunctional β-xylosidase/α-arabinosidases and one GH11 xylanase from *Humicola insolens* and their synergy in the degradation of xylan. *Food Chem.* **148**, 381–387.

Zaldivar, J., Nielsen, J., & Olsson, L. (2001). Fuel ethanol production from lignocellulose: a challenge for metabolic engineering and process integration. *Appl. Microbiol. Biotechnol.* **56**, 17–34.

Zechel, D.L., & Withers, S.G. (2001). Dissection of nucleophilic and acid-base catalysis in glycosidases. *Curr. Opin. Chem. Biol.* **5**, 643–649.

Zeigler, D.R. (2014). The *Geobacillus* paradox: why is a thermophilic bacterial genus so prevalent on a mesophilic planet? *Microbiology* **160**, 1–11.

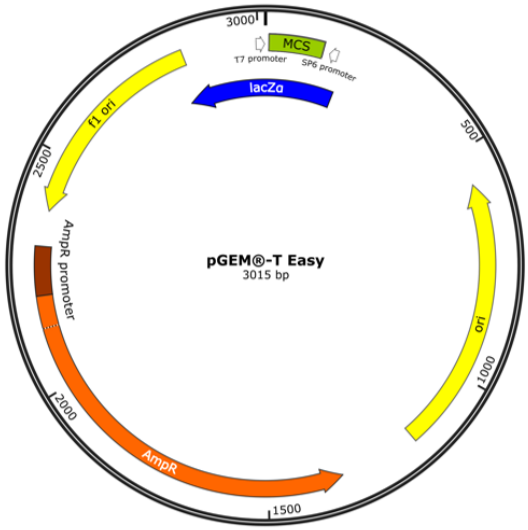
Zhang, Z.G., Yi, Z.L., Pei, X.Q., & Wu, Z.L. (2010). Improving the thermostability of Geobacillus stearothermophilus xylanase XT6 by directed evolution and site-directed mutagenesis. *Bioresour. Technol.* **101**, 9272–9278.

Zolotnitsky, G., Cogan, U., Adir, N., Solomon, V., Shoham, G., & Shoham, Y. (2004). Mapping glycoside hydrolase substrate subsites by isothermal titration calorimetry. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 11275–11280.

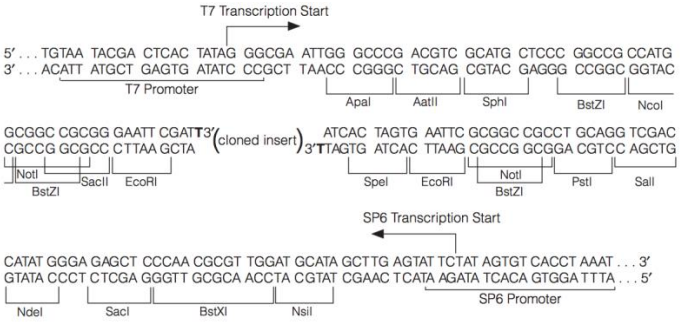
APPENDICES

I. Vector Maps

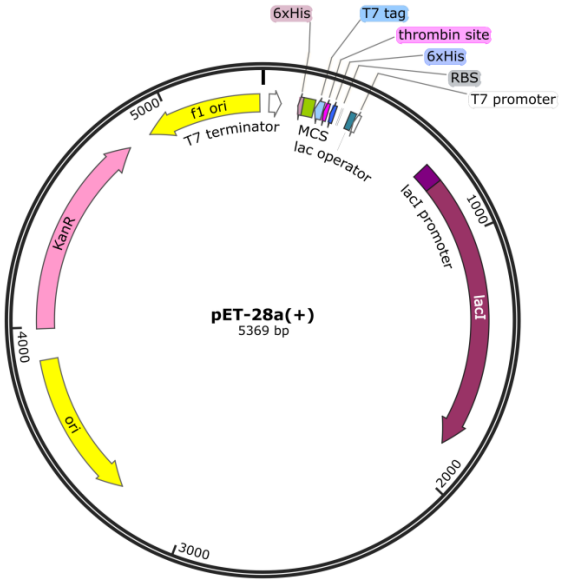
a. pGEM®-T Easy:



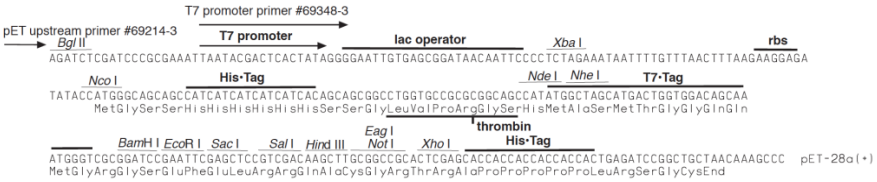
MCS:



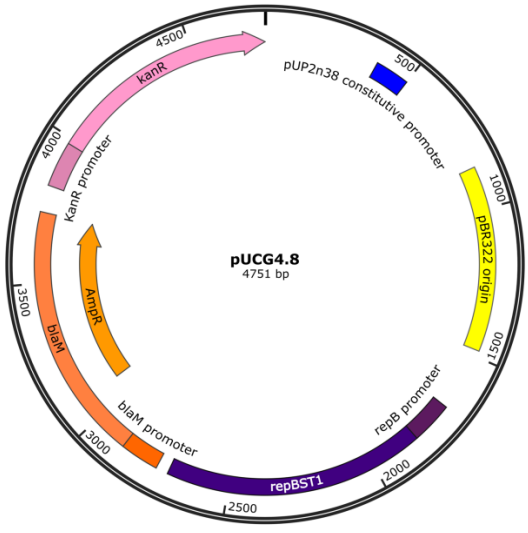
b. pET-28a(+)



MCS:



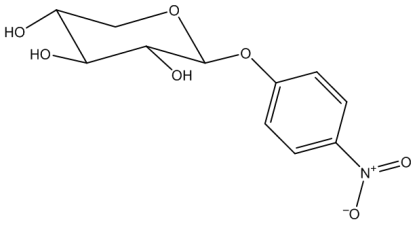
c. pUCG4.8



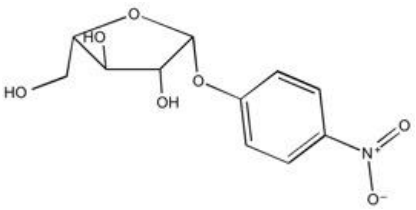
II. Structures of the chromogenic substrates tested

p-Nitrophenyl linked glycosides

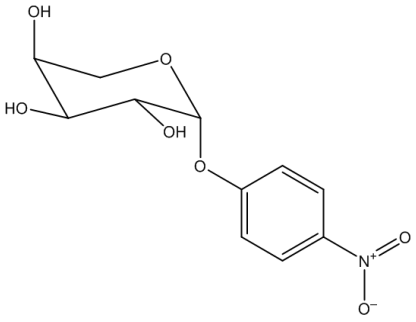
a. ***p*-Nitrophenyl-β-D-xylopyranoside**
(*p*NP-XP)



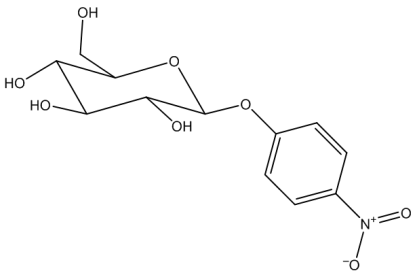
b. ***p*-Nitrophenyl-α-L-arabinofuranoside**
(*p*NP-AF)



c. ***p*-Nitrophenyl-α-L-arabinopyranoside**
(*p*NP-AP)



d. ***p*-Nitrophenyl-β-D-glucopyranoside**
(*p*NP-GP)



e. ***p*-Nitrophenyl N-acetyl-β-D-glucosaminide**
(*p*NP-NAG)